Adaptation to transposable elements in *Drosophila simulans*

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy

by Volha Paulouskaya

March 2019

TABLE OF CONTENTS

Abstract

Transposable elements (TEs) are genomic parasites that proliferate within host genomes, and can also invade new species. The *P*-element, a DNA-based transposable element, recently invaded two *Drosophila* species: *D. melanogaster* in the 20th century, and *D. simulans*, in the 21st. In both species, lines collected before the invasion are susceptible to 'hybrid dysgenesis', a syndrome of abnormal phenotypes that are due to *P*-element inflicted DNA damage. In *D. melanogaster*, lines collected after the invasion have evolved a maternally acting mechanism that suppresses the effects of the *P*-element and therefore hybrid dysgenesis. Extensive work has shown that PIWI-interacting small RNAs (piRNAs) are a key factor suppressing *P*-element induced hybrid dysgenesis. However, most of these studies were performed using lines collected many generations after the initial *P*-element invasion. In this thesis, I study lines of *D. simulans* collected early and late in the invasion of the *P*-element in that species. Similar to *D. melanogaster*, late in the invasion *D. simulans* shows abundant *P*-element derived piRNAs. Lines collected early in the invasion show substantial variation tolerance to the *P*-element. Surprisingly, however, these lines show no correlation between tolerance to *P*-element damage and expression of maternal *P*-element piRNAs, or other known factors influencing hybrid dysgenesis, suggesting mechanisms contribute to *P*-element suppression prior to the evolution of piRNA suppression. In addition to that, I identify piRNA-producing loci, piRNA clusters, in *D. simulans*.

Abbreviations

EtOH	Ethanol
HD	Hybrid Dysgenesis
HMW	High Molecular Weight
kb	kilobase
kDa	kilodalton
LTR	Long Terminal Repeat
nt	nucleotide
PCR	Polymerase Chain Reaction
piRNA	PIWI-interacting RNA
QTL	Quantitative Trait Loci
rpm	Revolutions per minute
RNAi	RNA interference
RT	Room Temperature
TAS	Telomere Associated Sequence
TE	Transposable element
TIR	Terminal Inverted Repeat

Chapter 1 General Introduction

1.1 TRANSPOSABLE ELEMENTS

One of the features of living organisms is the ability to reproduce, which involves transmitting genetic information to the next generation. The source of genetic information in living cells is DNA, and its most important functions involve providing cells with the information for molecular synthesis, and transmitting this information to progeny. Most genes make a positive contribution to an organism's fitness, are beneficial for reproduction and maintenance of the organism, and are transmitted to the next generation in a Mendelian fashion. Selection acts indirectly on the alleles of these genes, in proportion to their contribution to organismal fitness. However, other genes can spread in populations without being beneficial to organisms. Instead they use mechanisms to favour their own transmission, and are passed on to the next generation in a biased manner, more often than expected via Mendelian transmission. Genes with this kind of inheritance strategy are generally termed 'selfish genetic elements. One widespread example of selfish genetic elements is transposable elements, which have successfully invaded and spread in all eukaryotic species that have been investigated (Gregory, 2005; Feschotte and Pritham, 2007; Craig *et al.*, 2015). Transposable elements can be a major part of the genome, with human genome consisting of ~69% of TE sequences and some plant genomes up to 95% (Kronmiller and Wise, 2008; de Koning *et al.*, 2011). These 'jumping genes' are able to transpose, or change their location within and between

genomes, which normally results in an increase in their copy number and favours their transmission.

Transposable elements play important roles in genome evolution, regulation of gene expression and disease (Ayarpadikannan and Kim, 2014), therefore it is important to understand mechanisms and factors influencing their transmission.

1.1.1 Types of transposable elements

There are several major, independently evolved, types of TEs, defined by their transposition mechanisms. DNA transposable elements consist of 'cut-and-paste' transposons, *Helitrons*, and *Polintons* (Feschotte and Pritham, 2007; Craig *et al.*, 2015). 'Cut-and-paste' DNA transposons are flanked by terminal inverted repeats (TIRs) on both sides of the element, and transpose by a non-replicative mechanism, cutting themselves from one location in the genome, and inserting elsewhere in the genome. These transposons increase in copy number by recombination repair — transposition happens after DNA replication, the excised element gets inserted into a new place, and the gap left behind gets repaired using a sister chromatid containing a TE as a template (Engels *et al.*, 1990; Plasterk, 1991; Plasterk and Groenen, 1992; Hagemann and Craig, 1993; Nassif *et al.*, 1994; Arca *et al.*, 1997). *Helitrons* transpose via 'rolling circle mechanism', in which one strand of the DNA relocates to another position of the genome, where it serves as a template for DNA synthesis (Mendiola *et al.*, 1994; Kapitonov and Jurka, 2001). Complete copies of elements consist of two domains: the Rep domain involved in DNA transfer, cleavage and ligation, and the Hel domain that is involved in strand separation (Kapitonov and Jurka, 2001, 2007).

For *Polintons*, the mechanism of transposition of is not yet well understood (Kapitonov and Jurka, 2006; Pritham *et al.*, 2007).

In primate genomes, there appear to be no active DNA transposable elements, though inactive remnants remain (Pace and Feschotte, 2007). There are ~300,000 copies of these inactive elements, comprising about 2-3% of the human genome (Smit, 1999; Hattori *et al.*, 2000). The number of DNA transposons in the human genome is ~40 times larger than in the *D. melanogaster* genome (Feschotte and Pritham, 2007).

Some transposable elements transpose using an RNA intermediate. The second type of transposable elements are long terminal repeat (LTR) retrotransposons, which are more complex than DNA transposons — they transpose via an RNA intermediate, usually encode three enzymes and two structural proteins required for transposition, and are similar in structure to retroviruses (Havecker *et al.*, 2004). After transcription of the element, some mRNA is used for as a template for translation of 5 or 6 proteins encoded by the element. Alternatively, the RNA transcript may be encapsulated using those 5 or 6 proteins and then reverse transcribed to cDNA, which is inserted is a new genomic location (Coffin *et al.*, 1997).

The third class of transposable elements is the long interspersed nuclear elements (LINEs, or non-LTR retrotransposons) that, like LTRs, use an RNA intermediate in transposition. Autonomous elements of this type encode one or two multifunctional proteins — reverse transcriptase for synthesising cDNA from RNA and endonuclease for generating single-stranded breaks in the genomic DNA at the place of the insertion. TEs that transpose via RNA intermediate, LTR and non-LTR

retrotransposons, increase their copy number at each transposition event as excision of the element does not happen in this type of transposition (Luan *et al.*, 1993; Boeke and Stoye, 1997).

1.1.2 Transposable elements and regulation of their activity

Transposable elements are very successful in spreading within and between species: all three classes of TEs have successfully invaded and spread though virtually all of the species that have been investigated so far, and most species have several copies of different types of TEs in their genomes (Craig *et al.*, 2015).

Transposable element activity is a source of mutations. As with any other kind of mutation, transposition events often have deleterious effects (Burt and Trivers, 2006). Active TEs can disrupt protein coding genes by inserting themselves into coding regions, and can cause chromosomal breakage, genome rearrangements and ectopic recombination (Montgomery *et al.*, 1987; Brookfield, 1991). The deleterious effects the TEs have forced animals to evolve to protect their genomes, both in somatic tissues and in the germ line using a variety of mechanisms, including chromatin and DNA modifications, RNA interference (Slotkin and Martienssen, 2007).

Chromatin modifications — DNA methylation and histone modifications — suppress TEs transcription in different organisms. In *Arabidopsis,* methylation of cytosine residues in CHH motifs in the genome silences transposable elements, resulting in DNA being hypermethylated in the regions of silent TEs (Zhang *et al.*, 2006; Cokus *et al.*, 2008). DNA methylation is not present in all of the eukaryotic

organisms: *Drosophila* seem to not have a gene encoding the enzyme for methylating cytosine residues (Dunwell *et al.*, 2013). However, silencing modifications of chromatin, similar to the one found in other organisms, are present in *Drosophila* (Martens *et al.*, 2005). For example, histone H3 methylated at lysine 9, a sign of transcriptionally inactive chromatin, is elevated in the TE-rich regions of the genome. Mutations in the genes needed for histone tail methylation results in TE derepression (Martens *et al.*, 2005).

RNA interference (RNAi) is another way to control TE activity. Double-stranded RNAs (dsRNAs) homologous to transposable elements are cleaved into small interfering RNAs (siRNAs) usually 21-22 nt long by Dicer-family proteins. siRNAs are loaded into the RNA-induced silencing complex (RISC), and these protein-RNA complexes recognise and degrade complementary transcripts (Slotkin and Martienssen, 2007). In *A. thaliana*, RNAi is involved in RNA-dependent DNA methylation and TE suppression. siRNAs of two classes have been discovered in this species — smaller class of 21-22 nt involved in RNAi and larger class of 24-26 nt that are derived mostly from transposable elements and take part in RNA-dependent DNA methylation (Hamilton *et al.*, 2002; Qi *et al.*, 2006).

Transposable elements are mostly active in the germ line of animals, which makes this tissue sensitive to the consequences of TE activity. Along with other ways to suppress TEs, there is another process regulating TE activity in the germ line — PIWI-interacting RNAs (piRNAs), small non-coding RNAs that act against TEs and silence them. While all small RNAs are bound by the proteins of the same family, Argonaute (Lander *et al.,* 2001.) proteins, and their mechanisms of target cleavage

are similar, the process of formation of piRNAs differs from the other types of small RNAs (Carmell *et al.*, 2002).

1.1.3 piRNA clusters and piRNA biogenesis

piRNAs and their protein partners are expressed mainly in the germline cells, where transposable elements are highly active (Vagin *et al.*, 2006; Aravin *et al.*, 2007; Brennecke *et al.*, 2007; Houwing *et al.*, 2007; Aravin *et al.*, 2008; Kuramochi-Miyagawa *et al.*, 2008). piRNAs are longer compared to other small RNAs and range from 24 to 36 nt in length. Sequences of piRNAs are quite diverse and contain no particular motifs, except an enrichment of uracil at the 5' end (1U bias) (Brennecke *et al.*, 2007). The principal origins of piRNAs in the genome are several heterochromatic regions, which consist of nested, degenerate copies of transposable elements. These regions can be up more than 200 kb long and are called 'piRNA clusters' (Brennecke *et al.*, 2007).

piRNAs are predominantly expressed in gonads, where transposable elements are highly active. *Drosophila* oocytes consist of germline cells and somatic support cells. Both of these types of cells depend on the piRNA pathway to protect the genome against transposons. However, the mechanisms of silencing the transposon differs between somatic support cells and germline cells (Senti and Brennecke, 2010).

In flies, two types of piRNA clusters have been described: uni-strand, which are the main piRNA clusters expressed in the somatic support cells, and which contain majority of the transposable elements in the same orientation and are

transcribed from one strand, and more common dual-strand clusters, which consist of transposable elements inserted in both orientations and are transcribed from both strands (Brennecke *et al.*, 2007; Huang *et al.*, 2017). The common feature of the two cluster types is the presence of H3K9me3 (histone 3 lysine 9 tri-methylation), an epigenetic mark that is generally found in heterochromatic regions of the genome, and which silences transcription of the regions carrying it (Rangan *et al.*, 2011; Sienski *et al.*, 2012; Le Thomas *et al.*, 2013; Rozhkov *et al.*, 2013; Klenov *et al.*, 2014; Mohn *et al.*, 2014; Zhang *et al.*, 2014).

The two cluster types differ in transcription in two ways: first, whether one or both strands of the cluster is transcribed — one for uni-strand clusters (comparable to canonical mRNA transcription), or both for dual-strand clusters — and second, in transcription initiation — while uni-strand clusters have Pol II promoters, dual-strand clusters seem to not have established promoters, and transcription starts at multiple sites (Huang *et al.*, 2017).

Transcription of piRNA clusters gives rise to long piRNA precursors that are transported to the cytoplasm for further processing and generation of mature 23-36-nt 'primary' piRNAs (Le Thomas *et al.*, 2014). Mature piRNAs are bound by members of PIWI protein family – Piwi, Ago3 and Aub – and these ribonuclearprotein complexes (RNPs) target transposable element transcripts (Brennecke *et al.*, 2007).

While primary piRNAs are the main piRNAs expressed in the somatic support tissue, in the germline, primary piRNAs are also used to generate secondary piRNAs (Malone *et al.*, 2009). That is, in the germline cells, all three members of Piwi protein

family are expressed – Piwi, Aub and Ago3, whereas in gonadal somatic cells only Piwi is present. The fact that gonadal somatic cells lack Aub and Ago3 indicates that piRNA processing is different in those cells compared to gonadal germline cells (Malone *et al.*, 2009). In fact, somatic piRNA production relies on piRNA production from uni-strand cluster found at the distal end of X chromosome in *D. melanogaster*, *flamenco*, and the Piwi protein. In contrast, in the germline cells, piRNAs are mostly produced by dual-strand clusters and Piwi, Aub and Ago3 participate in piRNA processing during the so-called "ping-pong" cycle (Brennecke *et al.*, 2007; Malone *et al.*, 2009). First, piRNAs antisense to the TE mRNA are generated from a piRNA cluster. These piRNAs form a complex with one of the two proteins, Piwi or Aub, which recognize and cleave the sense transcript of the transposable elements, opposite 10 and 11 nt of the piRNA 5' end of the antisense piRNA. The resulting cleaved product is loaded onto Aub protein, which then recognises and cleaves piRNA cluster transcript, producing a new antisense small RNA, which is then used again to cut the sense transcript. The entire cycle repeats itself which leads to production of piRNAs against TEs that are actively transcribed, and degradation of the transposable element mRNA (Brennecke *et al.*, 2007; Senti and Brennecke, 2010) (Fig. 1.1).

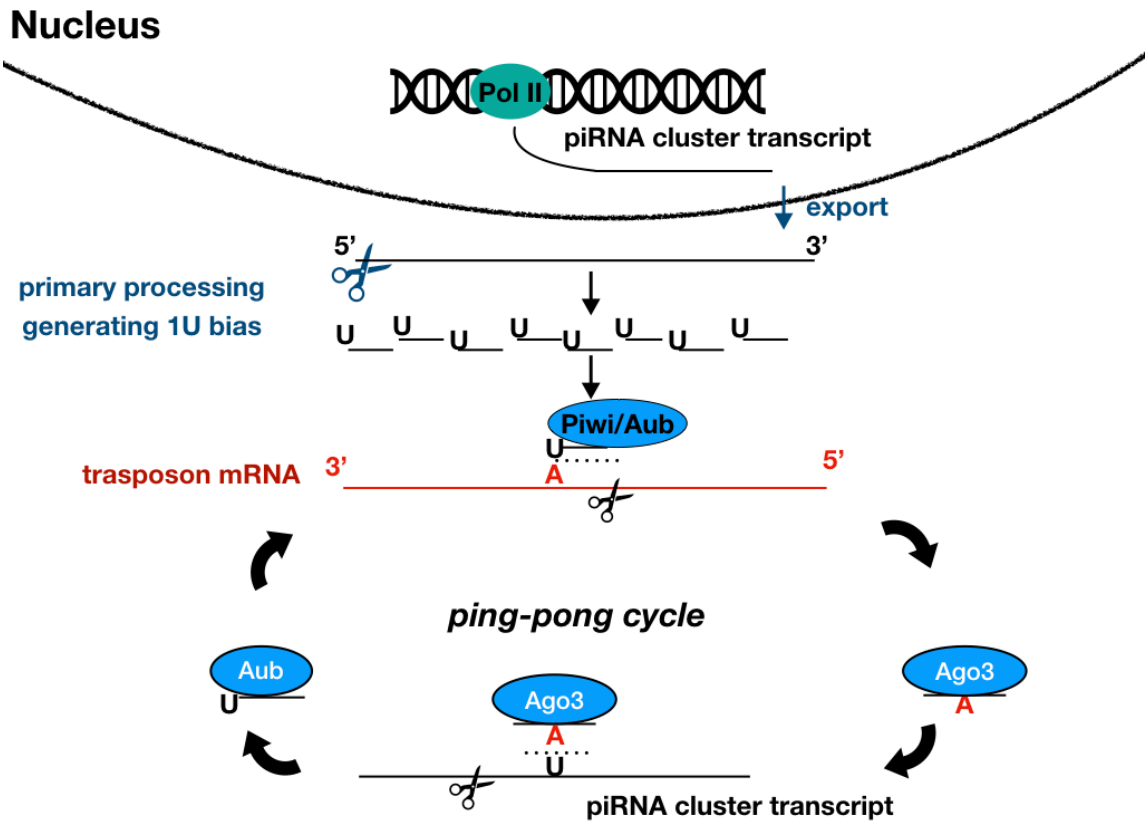


Figure 1.1. Illustrated piRNA ping-pong cycle. Following Brennecke *et al.*, 2007.

1.1.4 Domestication of transposable elements

Most of the mutations resulting from transposition will be deleterious, some will be neutral or beneficial, just as any other type of mutation. Moreover, transposition can be a source of unusual kinds of structural mutations – inversions, duplications and chromosomal rearrangements, which makes TEs a great potential source of evolutionary innovation (Burt and Trivers, 2006).

Transposable elements that are neutral to the host can be fixed in a population by genetic drift. In this case, the pattern of transposable element

sequence evolution is the same as of neutrally evolving sequences. These transposable element insertions accumulate mutations over time, and are often no longer able to transpose, leading to the accumulation of 'immobilised' TEs across the genome (Carr *et al.*, 2012; Wacholder *et al.*, 2014).

In some cases, insertions of the transposable elements will be co-opted by the host to perform cellular functions (Miller *et al.*, 1997). TE insertions can alter gene expression at transcriptional and posttranscriptional levels and there are many examples of TE sequence 'domestication' (Bejerano *et al.*, 2006; Cohen *et al.*, 2009; Rebollo *et al.*, 2012; Thompson *et al.*, 2016; Chuong *et al.*, 2017). Changing gene expression at the transcriptional level includes cases when a TE insertion disrupts regulatory sequences of the gene, provides an alternative binding site for a transcription factor or recruits heterochromatin formation factors therefore silencing the adjacent genes (Feschotte, 2008; Elbarbary *et al.*, 2016). At the posttranscriptional level, TE insertions within introns can interfere with a normal splicing pattern, and 3'UTR insertions may result in alternative polyadenylation signal (Feschotte, 2008; Elbarbary *et al.*, 2016). Recent studies have discovered that TE insertions can act as tissue-specific promoters and contribute to the pool of regulatory elements across the genome providing binding sites for transcription factors (Cohen *et al.*, 2009; Rebollo *et al.*, 2012; Thompson *et al.*, 2016).

There are several examples of TE insertions that are beneficial for their hosts. In tetrapods, an insertion of a SINE (non-LTR retroelement) acts as an enhancer during brain development (Bejerano *et al.*, 2006). Choung *et al* reported that endogenous retroviruses (ERV) regulate inflammatory response by functioning as

interferon-inducible enhancer (Chuong *et al.*, 2017). In *D. melanogaster,* a DNA transposon mediates an antioxidant response which provides increased resistance to oxidative stress (Guio *et al.*, 2014). These are only a few examples for a TE sequence being co-opted by the host to be functional non-coding elements, there are many more cases of TE being a regulatory element and it is likely that many more TE insertions playing a crucial role in gene regulation yet to be investigated (Bourque, 2009; Rebollo *et al.*, 2012; Kapusta and Feschotte, 2014).

Some other examples of TE domestication involve not only TE sequences being co-opted to serve host's function as non-coding elements, but also TE proteins being domesticated and now used by the cells (Jangam *et al.*, 2017). V(D)J recombination that is a part of adaptive immune response creates an infinite number of antibodies in B and T cell. Recombinases involved in this process, RAG1 and RAG2, were shown to have origins of DNA transposon *Transib* (Kapitonov and Jurka, 2005; Huang *et al.*, 2016; Carmona and Schatz, 2017). Another example of TE protein domestication may be CRISPR-Cas system, with CRISPR repeats sharing sequence similarity with TIRs of *Casposons* and Cas1 protein sharing similarity with *Casposons* transposases (Krupovic and Makarova, 2014; Hickman and Dyda, 2015; Béguin *et al.*, 2016). An example of retroviruses genes being domesticated because of the conflict between mother and embryo is *Env* gene, that gave rise to syncytins – proteins vital for placenta functions (Esnault *et al.*, 2013; Lavialle *et al.*, 2013; Cornelis *et al.*, 2014).

1.1.5 Life cycle of transposable elements

Transposable elements spread in genomes by producing copies of themselves, which can lead to the accumulation of hundreds of copies of the element in a single genome. Nevertheless, despite their activity and ability to increase in numbers, transposable elements can go extinct within host genomes. They can accumulate mutations and become no longer active, and therefore unable to transpose. In this case, a host genome contains only degenerated copies of an element. The human genome, for example, consists of hundreds of thousands inactive transposons (Lander *et al.*, 2001). This can result from parasitism of active elements by inactive elements. Inactive elements are not able to transpose by themselves, but they can use the transposition machinery produced by active elements and therefore move around the genome (Piskurek and Jackson, 2012).

There is a hypothesis that in order to be active over a long period of time and not to become extinct, a TE needs to invade new species (Burt and Trivers, 2006). Once a functional element invades a new species, its transposition rate is supposed to be quite high as the host cannot yet regulate its activity and gives a TE an opportunity to increase in frequency (Brookfield, 2005). This excessive transposition may be harmful to the host, as it results in DNA breakage, may lead to ectopic recombination, and may disrupt essential genes. Over time, the host factors can evolve to protect the genome from the deleterious effects of the transposition and silence the TE, which may eventually cause the TE go extinct in this species if it is eliminated faster than it transposes to make new copies (Blumenstiel, 2010). Alternatively, a transposable element can be “domesticated” by the host and serve

as regulatory elements across the genome. On average, a transposable element can persist over time if it invades another species at least once before going extinct in the host species.

One of the best studied examples of a horizontally transmitted DNA-based transposon is the *P*-element, a transposable element that has recently invaded and spread in *D. melanogaster* within 50 years (Kidwell *et al.*, 1977; Kidwell, 1985; Anxolabehere *et al.*, 1988). The characterisation of *P*-element's activity has made it a powerful genetic tool and has resulted in better understanding of the mobilisation of DNA-based transposons (Cooley *et al.*, 1988; Bachmann and Knust, 2008; Hummel and Klambt, 2008).

1.2 *P*-ELEMENT

1.2.1 Hybrid dysgenesis

The *P*-element, a DNA based transposable element, was first discovered in *D. melanogaster* in early 1970s, when laboratory strains were crossed to strains recently isolated from natural populations (Hiraizumi, 1971; Hiraizumi *et al.*, 1973). The offspring of such crosses demonstrated high mutation rates, chromosomal rearrangements and male recombination (which is not normal for *Drosophila*), high rates of sterility and abnormally small gonads (Hiraizumi, 1971; Hiraizumi *et al.*, 1973; Kidwell and Kidwell, 1975; Kidwell *et al.*, 1977; Engels and Preston, 1979; Schaefer *et al.*, 1979). This phenomenon, called 'hybrid dysgenesis', is observed at 29°C when males from a P strain – a paternally contributing strain with multiple *P*-elements in the genome – are crossed to females from M type strains – maternally

contributing strains that are devoid of *P*-elements. The reciprocal cross, when a M type male devoid of *P*-elements is crossed to a carrying *P*-elements P type female, results in normal, 'non-dysgenic' offspring (Kidwell and Kidwell, 1975; Kidwell *et al.*, 1977). Hybrid dysgenesis does not occur when both crossed lines are of the same type, as in P male x P female or M male x M female. Flies that cannot regulate *P*-element activity are thought of as M cytotype; they are permissive for hybrid dysgenesis and do not have *P*-elements in their genome. Flies that can repress *P*-element transposition in the germ line, are restrictive to hybrid dysgenesis and can induce it when crossed to M cytotype female, they are considered to have P cytotype (Bingham *et al.*, 1982; Anxolabehere *et al.*, 1985; Anxolabehere *et al.*, 1988). In addition to M and P cytotypes, there are also other cytotypes — M' and Q. M' type flies normally have some deleted copies of the *P*-element but are not able to induce or repress hybrid dysgenesis, whereas Q types, that also have some parts of the *P*-element DNA, can repress hybrid dysgenesis but not induce it (Bingham *et al.*, 1982; Anxolabehere *et al.*, 1985; Anxolabehere *et al.*, 1988; Itoh *et al.*, 2001; Itoh *et al.*, 2004; Ogura *et al.*, 2007; Fukui *et al.*, 2008). The activity of *P*-elements in the germline of dysgenic offspring was later discovered to be the cause of hybrid dysgenesis (Bingham *et al.*, 1982; Eggleston *et al.*, 1988; Lemaitre *et al.*, 1993; Khurana *et al.*, 2011).

1.2.2 Structure of the *P*-element and regulation of its activity

The *P*-element is one of the best studied examples of DNA-based transposable elements. Full-length *P*-elements are 2907 bp in size and both ends of the elements

are flanked by 31 bp terminal inverted repeats (TIRs) (O'Hare and Rubin, 1983). The element consists of four exons, that can be alternatively spliced and produce two proteins – a transposase and a repressor of transposition. When all of the introns are spliced out, an active 87-kDa transposase is translated; retention of the third intron leads to a premature STOP codon and expression of a smaller 66-kDa protein that acts as a repressor of transposition (Laski *et al.*, 1986; Rio *et al.*, 1986).

The *P*-element transposes via 'cut-and-paste' mechanism and increases in copy number through recombination repair mechanism during replication. Full-length elements encode transposase and can transpose autonomously. However, only about 1/3 of all the *P*-elements found in *D. melanogaster* are full-length, the rest are internally deleted copies between 0.5 kb to 2.9 kb in size that can transpose only in the presence of full-length elements by using their transposases (O'Hare and Rubin, 1983).

P-elements are mostly active in germline cells, and this germline specificity is achieved by alternative splicing of the transcript; the third intron is retained in the soma and a repressor of transposition is expressed (Laski *et al.*, 1986; Rio *et al.*, 1986). Moreover, other *P*-element derived sequences can serve as repressors of transposition. In contrast to the 66-kDa repressor, the expression of other repressors is not regulated by alternative splicing. These repressor proteins are classified into 2 types. Type I repressors, including 66-kDa protein, do not have any deletions in 2/3 of the 5' end of the *P*-element (Gloor *et al.*, 1993). Type II repressors are shorter in sequence and mainly have first and part of the second exon of the *P*-element (Andrews and Gloor, 1995). One well known type II suppressor is the KP

element, which codes for a 207 amino acid polypeptide and is derived from a *P*-element by a deletion of nucleotides 808-2560 (Black *et al.*, 1987). Both type I and type II repressors suppress *P*-element transposition in somatic and germline cells. *P*-element induced hybrid dysgenesis (Pasyukova *et al.*, 2004) happens when an M type female, devoid of *P*-elements, is crossed to a P type male. Therefore, there is a maternally contributing factor protecting the offspring from negative consequences of hybrid dysgenesis. None of the type I and II repressors acts as maternal repressors (Gloor *et al.*, 1993; Andrews and Gloor, 1995). Classic work hinted at the genetic basis of the regulation of the *P*-element activity. A study of a sample of flies from natural populations of *D. melanogaster* revealed excess accumulation of the *P*-element at the tip of the X chromosome, in the telomere-associated sequence (X-TAS) (Ajioka and Eanes, 1989). Later work has shown that transgenes inserted into X-TAS cause silencing of the same transgenes found in euchromatic regions (Todeschini *et al.*, 2010). Insertions into this region are probably beneficial because they suppress the *P*-element. Now, the tip of the X chromosome is known to be a piRNA-producing locus, and piRNAs are known to be a maternally acting factor that protects the germline from negative consequences of *P*-element activity (Brennecke *et al.*, 2007; Senti and Brennecke, 2010).

Components of the piRNA pathway seem to not act to lower the overall expression of the *P*-element transcript. Instead, they act to reduce the levels of spliced, transposon-coding transcripts. Comparison of *P*-element expression levels between genetically identical 'dysgenic' and 'non-dysgenic' crosses show ~10 fold higher expression of the active transposon-coding version of the transcript, whereas

the overall levels of *P*-element transcripts seem to differ ~2 fold (Teixeira *et al.*, 2017). The exact mechanism of regulation of the splicing by piRNA pathway components yet remains unclear.

1.2.3 *P*-elements in *D. simulans*

Transposable elements are not only transmitted vertically, they are also transmitted horizontally by invading new species. Recently, it was found that the *P*-element, which invaded *D. melanogaster* in the 20th century, has now invaded and spread through *D. simulans* several times faster (Anxolabehere *et al.*, 1988; Kofler *et al.*, 2015; Hill *et al.*, 2016). The *P*-element was most likely transmitted to *D. simulans* from *D. melanogaster*: There is only a single base pair difference between *P*-elements of *D. simulans* and *D. melanogaster* (2040 G>A). The variant that is fixed in *D. simulans* populations segregates at low frequency in *D. melanogaster* (0.16-2%), suggesting that the *P*-element transmission to *D. simulans* was a single event (Kofler *et al.*, 2015). It seems that transmission was horizontal as hybrids of these two species are normally inviable or infertile (Sturtevant, 1920; Lachaise *et al.*, 1986).

An interesting feature of the spread of the *P*-element in *D. simulans* is that, for the first time, the entire process of invasion of the transposable element has been captured. This gives a unique opportunity to study the process of adaptation of the host defence system to the new TE. In both species, *D. melanogaster* and *D. simulans*, flies collected before the invasion cannot regulate *P*-element activity, while ones collected at the end of the invasion can, as shown by their resistance to *P*-element-induced "hybrid dysgenesis" (Kidwell and Novy, 1979; Hill *et al.*, 2016). As in *D.*

melanogaster, there are three major cytotypes in *D. simulans*: M, P and Q (Hill *et al.*, 2016). Flies that have Q cytotype, as in *D. melanogaster*, are able to repress hybrid dysgenesis but not induce it. There is variation in the ability of Q cytotype flies to supress HD (Hill *et al.*, 2016). Recent work on *D. melanogaster* revealed genetic basis of the tolerance to the *P*-element activity in flies lacking *P*-elements. The *bruno* locus on chromosome 2L is a strong candidate for the origin of P tolerance due to its role in cystoblast differentiation (Parisi *et al.*, 2001; Wang and Lin, 2007), as the genomic region that contains it explains ~35% of variation in P tolerance (Kelleher *et al.*, 2018).

1.3 Aims

Lines of *D. simulans* collected during an early phase of *P*-element invasion show variation in tolerance to *P*-element induced hybrid dysgenesis. In this thesis, I describe tolerance to the *P*-element-induced hybrid dysgenesis early in the *D. simulans* invasion, and mechanisms that might or might not be involved. I perform crosses to genetically characterise the tolerance in these populations. Finally, I identify piRNA-producing loci in *D. simulans*, and compare the architecture of some of the piRNA clusters between two *D. simulans* and *D. melanogaster*.

Chapter 2

2.1 INTRODUCTION

Most genes have important functions for the viability of organisms. However, some genes are selfish, not necessarily contributing to the organismal fitness, and may even have deleterious consequences for their hosts (Burt and Trivers, 2006). Nevertheless, they can persist, due to mechanisms that manipulate transmission to the next generation in their favour (Werren *et al.*, 1988). The most taxonomically widespread example of a selfish genetic element is transposable elements (Arkhipova and Meselson, 2000; Wicker *et al.*, 2007). Transposable elements are a type of genetic parasite that changes location within genomes ('transposes'), copying themselves from one location in to another and thereby increasing their copy number within the genome. TEs are widespread and have been found in all eukaryotic species investigated so far (Gregory, 2005), they account for ~ 60% of the human genome and up to 95% of some plant genome (Kronmiller and Wise, 2008). TEs proliferate both by spreading within genomes, and by invading new species (Engels, 1992).

As with any other mutation, effects of transposition on the host are mostly deleterious, as transposition results in DNA breakage, can lead to ectopic recombination, and can disrupt essential genes (Langley *et al.*, 1988; Hua-Van *et al.*, 2011; Chuong *et al.*, 2017). In response to these deleterious effects, organisms have evolved a variety of mechanisms to protect their genomes, both in somatic tissues

and in the germ line (Slotkin and Martienssen, 2007), including chromatin modifications, RNAi and piRNAs (Brennecke *et al.*, 2007; Ozata *et al.*, 2019).

One of the best studied examples of a transposable element is the *P*-element, a DNA-based transposon which invaded and spread worldwide in *D. melanogaster* over several decades in the 20th century (Bingham *et al.*, 1982; Anxolabehere *et al.*, 1988) and which invaded and spread even faster in *D. simulans* (Kofler *et al.*, 2015; Hill *et al*., 2016). The *P*-element can be highly costly: it can induce hybrid dysgenesis (Pasyukova *et al.*, 2004), a syndrome consisting of high mutation rates, chromosomal rearrangements, and atypically small gonads which may be sterile (Kidwell *et al.*, 1977; Kidwell and Novy, 1979). HD occurs in the offspring of 'dysgenic' crosses, when males carrying *P*-elements in its genome — 'P type' males — are crossed to the females that lack them – 'M type' females (Kidwell *et al.*, 1977). The *P*-element encodes a single protein coding gene, transposase, an enzyme required for transposition (Kidwell and Novy, 1979; Misra and Rio, 1990). HD is likely a by-product of uncontrolled transposition in the germ line, which involves double-stranded DNA breaks catalysed by the endonuclease activity of transposase (Beall and Rio, 1996; McVey *et al.*, 2004). M type females are unable to regulate *P*-element transposition in their offspring, and the resulting unsuppressed transposition can yield DNA damage sufficient to trigger programmed cell death in the developing germ-line, resulting in malformed, 'dysgenic' gonads (Kidwell *et al.*, 1977). The substantial costs imposed by uncontrolled transposition suggest that species should rapidly evolve mechanisms to suppress TEs.

Indeed, organisms have evolved several ways to suppress TE activity, including that of the *P*-element. In fact, most wild lines of *D. melanogaster* are neither P type nor M type – they do not have full length *P*-elements in their genomes and therefore do not induce HD but are *tolerant* to the *P*-element-induced HD and do not show negative phenotypic consequences of the *P*-element transposition (Kidwell and Novy, 1979). These 'not P type' and 'not M' type lines are considered to be 'Q type' lines (Kidwell, 1985). Given that Q-type lines are predominant in wild populations, understanding how they suppress *P*-element activity is a critical, but understudied aspect of understanding the dynamics of transposable elements in natural populations. The failure of M type females to regulate transposition in the germ line of their offspring appears to be due to an absence of PIWI-interacting RNAs ('piRNAs'), which are small non-coding PIWI-interacting RNAs that act to silence TEs in the germ-line of most animals (Brennecke *et al.*, 2007). These 'piRNAs' are encoded by sequences homologous to TEs which are concentrated into piRNA clusters, usually in heterochromatic regions of the genome (Brennecke *et al.*, 2007). In flies, piRNAs are loaded into the egg by the female parent, and HD results from a lack of piRNAs homologous to a TE carried by the male parent. The offspring of crosses between M type females and P type males inherit paternal DNA that codes for piRNAs that suppress *P*-elements. However, early in life expression of these piRNAs is insufficient to protect the developing germ-line (Khurana *et al.*, 2011).

In *D. melanogaster,* the piRNA pathway regulates TEs in at least three ways: by degrading TE transcripts *via* cleavage by Argonaute proteins in the cytoplasm, as is typical for small RNAs (Brennecke *et al.*, 2007), by inducing chromatin state

changes that transcriptionally silence TEs in the nucleus, and, at least in the case of the *P*-element, by suppressing splicing so that no functional transposase is made (Klenov *et al.*, 2007; Senti and Brennecke, 2010; Klenov *et al.*, 2011; Wang and Elgin, 2011; Le Thomas *et al.*, 2013; Teixeira *et al.*, 2017).Therefore, piRNAs are a vital defence against transposable elements in the germline, where TEs are mostly active.

However, to date work on mechanisms of TE suppression has focused on TEs that have been present in species for a substantial number of generations, and are fixed in these species. Here, I examine tolerance to *P*-element induced hybrid dysgenesis in *D. simulans* collected during the early phases of the invasion. Specifically, I test 28 isofemale lines for maternal suppression of female gonadal dysgenesis – tolerance to *P*-element-induced HD, and find substantial variation for this phenotype. Surprisingly, I am unable to find any association between maternal suppression and maternal piRNA production for either of the dysgenic elements present. These results suggest some other mechanism must exist that either suppresses the TEs, or allows tolerance of their effects.

2.2 Materials and Methods

2.2.1 Fly strains

The *D. simulans* isofemale lines used in this study were collected in 2009 in Athens and Morben, Georgia, USA (by P. Haddrill and A. Paaby), and maintained on standard cornmeal-molasses-yeast-agar *Drosophila* medium. For a tester P line I used Cro18, an isofemale line collected in 2014 and known to induce hybrid dysgenesis in susceptible flies.

2.2.2 Assaying P tolerance across the fly isofemale lines

I first assayed the 28 isofemale lines, 4 P type and 24 Q (not able to induce HD, but are tolerant to it) type, for their ability to suppress *P*-element induced ovarian dysgenesis. The 28 isofemale lines of *D. simulans* flies were reciprocally crossed at 29°C, the temperature at which HD can be seen, to the tester P Cro18 isofemale line. For each cross, I used 5 virgin females and males from each line and flies were left for a total of 8 days to lay eggs. I dissected 30 three to four days old F1 females from each cross and checked them for the presence or absence of two well-formed ovaries. I considered the females that lacked two normal ovaries to be dysgenic, indicating that that individual failed to suppress apoptosis induced by the *P*-element activity. I tested for significant differences in the proportion of dysgenic females between reciprocals using a Fisher's exact test.

2.2.3 Choosing lines for in depth study

I then chose twelve lines of flies for sequencing of piRNAs. The lines were chosen based on the proportion of hybrid dysgenesis their F1 offspring showed – ranging from the least to the most tolerant. As potential differences in the *P*-element copy number in the tester line may cause differences in the strength of hybrid dysgenesis, I reduced potential variation in copy number within the tester line by single pair sib-mating. Each of three sublines was set up by crossing one virgin female of Cro18 to one male of the same line, and all three sublines were used to test the chosen twelve isofemale lines for hybrid dysgenesis in replicate. All female F1 offspring (mean=103) of these crosses were tested for hybrid dysgenesis. As the results were consistent with our initial assay of P tolerance across the isofemale lines (Hill et al., 2016), I used these 12 lines for in depth study of TE tolerance. I also checked for the presence of the copies of the full length *P*-element in the genome of the chosen isofemale lines by PCR and confirmed none of the studied isofemale lines had a full length *P*-elements.

2.2.4 Sequencing *P*-elements in the studied strains

I extracted DNA from the chosen 12 lines using Qiagen DNeasy Blood & Tissue Kit. I used *P*-element specific primers for each of the four *P*-element exons (Hill *et al.*, 2016), performed PCRs and sequenced PCR products. Primer sequences are listed in Supplementary Table 2.1.

2.2.5 qPCR for *hobo* copy number

D. simulans can carry a second HD inducing TE, called *hobo*. As a large *hobo* load might also drive HD/TE defences in an isoline, I need to know whether *hobo* correlates with HD susceptibility. To estimate the copy-number of the *hobo* elements in the genomes of the tested lines, I extracted DNA from 10 females per line in three biological replicates using Qiagen DNeasy Blood & Tissue Kit. I performed qPCR using *hobo*-specific primers (amplified region 1232-1402, product size 170 bp) and a RP49 reference gene (Supplementary Table 2.2).

2.2.6 Total RNA preparation

I isolated ovaries from ten 3-4 days old females from each of the 12 lines, homogenized in Trizol (Invitrogen) and frozen in liquid nitrogen. I extracted total RNA according to manufacturer's instructions, using 5PRIME heavy Phase Lock Gel tubes, measured RNA concentration using Nanodrop, and the quality of the samples by running 1 ug of the samples on a denaturating agarose gel. The final quantification was performed using Agilent Bioanalyser.

2.2.7 Small RNA sequencing

Library preparation (including depletion of the similarly sized 2S RNA) and sequencing was performed at Fasteris, with two biological replicates (consisting of 10 pairs of ovaries each) per line. After initial quality control, small RNAs were size

selected on acrylamide gel. Then the libraries were prepared as follows: single strand ligation of a 3' adapter, a *Drosophila* specific anti 2S RNA depletion, ligation of a 5' adapter, cDNA synthesis, and PCR amplification. The libraries were sequenced on an Illumina HiSeq 4000 1x50 lane and the data de-multiplexed according to the indexes and fastq files produced.

2.2.8 piRNA expression analysis

First, I removed 3'-adapters from the raw sequencing reads using Cutadapt software v1.10 (Martin, 2011). I also removed the reads shorter than 5 bp after trimming. I mapped the remaining reads to a database of *Drosophila* transposable element annotation v. 9.42 (available from http://www.flybase.org) allowing for 3 mismatches (-i 2 -l 40 -n 3 -M 1) and 6 mismatches (-i 2 -l 40 -n 6 -M 1) using *bwa aln* v 0.7.13 (Li and Durbin, 2009). I examined these reads for the 'ping-pong' signal characteristic of processed piRNAs (a 10 nt overlap between sense and anti-sense reads, and a U-bias at the 3'end of sense reads (Brennecke *et al.*, 2007)). To this end, I calculated the overlap between sense and anti-sense reads for overlap sizes ranging from 1 to 20 nt for each transposable element separately, and tested for an excess 10nt overlaps using a chi-square test. After mapping, I removed reads mapped with insertions and deletions. To select for piRNA reads, only reads >24 nucleotides were considered for further analysis. *P*-element coverage was calculated using *samtools* v 1.3.1 (Li *et al.*, 2009).

After this preliminary analysis, I mapped the reads to *D. simulans* miRNAs obtained from Flybase (ftp://ftp.flybase.net, dsim_r2.02_FB2017_04), allowing for 1 mismatch, again using *bwa aln*. Roughly 10% of the raw reads mapped to miRNAs.

Reads that mapped to miRNAs were removed from further analysis.

2.2.9 Differential expression analysis

After mapping the reads to the transposable elements, I performed differential expression analysis using voom, implemented in the Bioconductor package in R (Liu *et al.*, 2015). Voom estimates the mean-variance relationship in the data and uses this to compute weights for each gene (TE family in this case), and normalise the data. It tests for differential expression analysis using a log-linear model.

2.2.10 Testing for differences in expression and splicing efficiency of the *P*-element

I chose four lines— two with high (Lps5 and SGA27) and two with low (SGA14 and SGA26) tolerance for the *P*-element — to investigate differences in overall expression of the *P*-element and for differences in splicing efficiency of the *P*-element transcript. I crossed each line to Cro18 in both directions and at three different temperatures (18°C, 25°C and 29°C), and tested three biological replicates. To perform RT-qPCR, I isolated ovaries from F1 offspring of each of the crosses, as described above. Samples were treated with DNase (ThermoFisher Scientific) and cDNA synthesized using the Transcriptor First cDNA synthesis kit (Roche). I performed qPCR using KAPPA SYBR Green (Sigma) on cDNA using rp49 as a reference gene and two sets of primers for the *P*-element. The *P*-element primers corresponded to exon 2 (to assess the overall expression of the *P*-element) and IVS3

(to assess the expression of the spliced version of the *P*-element). Primers used and qPCR conditions are listed in Supplementary Table 2.3; some primer sequences were kindly provided by Z. Zhang.

2.2.11 Testing small RNA expression in the presence of an active *P*-element

As the presence of an actively transcribed *P*-element may change the expression of the small RNAs, I also sequenced small RNAs from the F1 offspring of the two most and two least P tolerant lines after each of them (four in total) were crossed to a P type tester isofemale line. For each of the four lines, I crossed 5 females to 2 males of the Cro18 tester P isofemale line. I isolated ovaries from three day-old female offspring of these crosses, and then proceeded with RNA extraction, small RNA sequencing and analysis as described above.

2.2.12 Analysis of the expression of piRNAs inherited from a male parent

I also examined the F1 females for expression of piRNAs against the *P*-element inherited from their male parent. To do this, I mapped piRNAs to the *P*-element references only. I compared the normalised coverage of the *P*-element among the paternal P type line (Cro18), four of the studied Q type lines (Lps5, SGA27, SGA26 and SGA26), and the F1 female offspring of Cro18 males crossed to each of the Q-type lines.

2.2.13 Testing dominance of the P tolerance

To genetically characterize *P*-element tolerance, I crossed the two most P tolerant lines to the two least tolerant ones (a total of 4 reciprocal crosses) in three replicates. For each cross, I used 5 virgin females and males from each line and flies were left for a total of 8 days to lay eggs. I tested the offspring of these crosses for HD by crossing the female offspring to a tester P isofemale line, Cro18, as described above.

2.3 RESULTS

2.3.1 Variation in tolerance to *P*-element induced HD

The *P*-element has invaded and spread in *D. simulans* in just over a decade, with the element present in only some flies collected in 2004 and in all of them collected in 2014 (Hill *et al.*, 2016). I studied *D. simulans* lines collected in eastern North America during a relatively early phase of the invasion of the *P*-element (2009). Most of these lines were found to lack full-length, and therefore potentially active, *P*-elements, but many contain partial copies (Hill *et al.*, 2016). I confirmed these results by attempting to amplify each of the four exons of the *P*-element; of the 28 lines, four were P type lines, and had partial (26) or no *P*-elements (Supplementary table 2.4).

I assayed each of these 28 lines, testing for *P*-element induced hybrid dysgenesis, which occurs when a female from a susceptible line is crossed to a tester male from a P-type line. I used Fisher's exact test to compare 30 F1 females from this direction of the cross to those from the reciprocal cross; 21 of 28 lines showed significant differences in the number of dysgenic progeny between reciprocal crosses, indicating hybrid dysgenesis (Fig. 2.1A), consistent with previous assays on the same lines (Hill *et al.*, 2016).

Among non-dysgenic crosses, however, there was substantial variation in the strength of dysgenesis, as measured by the proportion of dysgenic offspring.

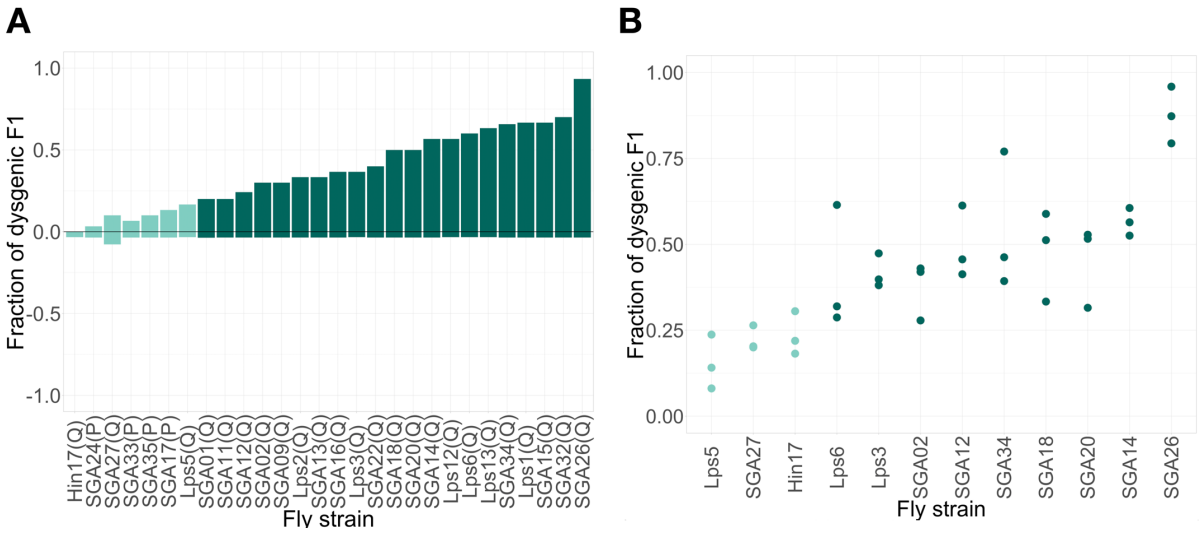


Figure 2.1. (**A**) The proportion of the F1 offspring (n=30) with hybrid dysgenesis when 32 lines of flies were crossed to a P type line. The top part of the graph represents the proportion of dysgenic offspring when a female from a susceptible line is crossed to a tester male from a P-type line; the bottom part – proportion of dysgenic offspring in a reciprocal cross. The bars are coloured according to the *p*-value of Fisher's exact test, with light green indicating $p > 0.05$ and dark green $p \leq 0.05$. (**B**) The proportion of the F1 offspring (mean number dissected = 103, range 31-173) with hybrid dysgenesis when females from 12 selected lines were crossed to males from the three sublines of the Cro18 tester P line.

To test whether lines show consistent variation in the tolerance to *P*-element induced HD, I selected a range of 12 Q type lines with different levels of susceptibility to HD – from most tolerant to least – and repeated the assay for HD in replicate. For this, I used three genetically homogenised sublines (see Methods), crossed each of the 12 Q lines to these three sublines (36 crosses in all), and dissected F1 female offspring to test for HD (mean number dissected = 103, range 31-173; Fig. 1B). I

tested for an effect of isofemale line, indicating heritable variation in tolerance to HD, using a binomial GLM. This test indicates significant variation in the proportion of the offspring of those crosses is due to line (GLM, N=36 *t*=3.7, *p*=4.43E-06).

2.3.2 Testing piRNA expression as a possible reason for P tolerance

As the main mechanism protecting the genome from TE activity is thought to involve piRNAs, one potential explanation variation in the degree of P tolerance could be due to differences in the piRNA expression among the studied lines. I therefore sequenced small RNAs from the ovaries of all 12 lines studied to test for differences in abundance of piRNAs homologous to the *P*-element. I obtained ~20 million reads per line (Supplementary Table 2.5). Approximately 40% of reads for each line mapped to the transposable element reference data base with 3 or fewer mismatches. Reads mapping to TEs had a bias for uridine at position 1 ("1U" bias) and ping-pong signatures characteristic of piRNAs (Brennecke *et al.*, 2007) (Supplementary figure 2.1).

I filtered reads mapping to *D. simulans* miRNAs (see Methods); ~10-11% of the reads in the size range typical of miRNAs (21-22 nt) mapped to these miRNAs, and I removed these reads from subsequent analysis. I mapped the remaining reads to the transposable elements of *Drosophila*, with ~40% of reads mapping, with 3 mismatches or fewer. I tested for expression differences using voom. First, I tested for differences between biological replicates of the same line, and did not find any significant differences (voom/limma t=-2.9, *p*=1.0). I then compared lines that were the most and least P tolerant, with the expectation that the lines most tolerant to hybrid dysgenesis would express more small RNAs mapping to the *P*-element than

lines with low P tolerance. Somewhat surprisingly, no significant expression differences were observed (voom/limma t = -4.00, p = 0.40) (Fig. 2.2).

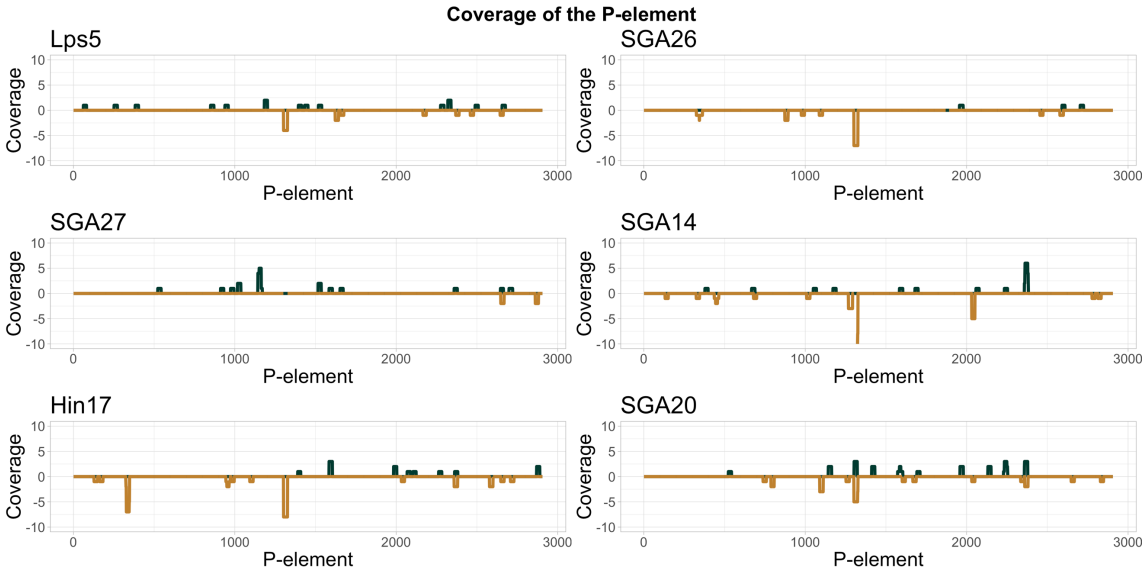


Figure 2.2. Normalised coverage of the *P*-element (mapped with 3 mismatches) in most tolerant (left: Lps5, SGA27, Hin17) and least tolerant lines (right: SGA26, SGA14, SGA20). Colours of the lines represent strandness of piRNAs (green: forward, oche: reverse). The coverage is based on the number of mapped reads out of ~20 million of raw reads obtained.

To ensure that these results did not depend on the particulars of arbitrary choices in the mapping parameters, I mapped all of the obtained reads allowing for 1, 3 and 6 mismatches (Brennecke *et al.*, 2007). Here, I mapped only to the *P*-element, allowing any reads that were primarily homologous to other elements to map to the *P*-element, in case piRNAs silencing other transposable elements can cross-react with the *P*-element transcript. Unexpectedly, I saw no significant differences in

coverage of the *P*-element between most and least P tolerant lines (voom/limma t=0.81, p=1.0).

2.3.3 Splicing of the *P*-element transcript in the studied lines

In addition to post-transcriptional silencing, *P*-element activity can also be regulated via splicing. In the soma, the retention of the third intron (IVS3) leads to a premature STOP codon and expression of a repressor of transposition (Rio *et al.*, 1986). In the germline of *D. melanogaster*, splicing suppression appears to be regulated by the piRNA pathway; while the *P*-element is expressed at the same level in reciprocal dysgenic and non-dysgenic crosses, in the dysgenic direction, there are higher levels of the active, spliced version of the *P*-element transcript than in the reciprocal cross (Teixeira *et al.*, 2017; Moon *et al.*, 2018).

Regardless of its mechanism, differences in tolerance to the *P*-element may be explained by differences in splicing regulation. To test for this, I crossed the two most and least P tolerant lines to a tester P line in both directions and at three different temperatures – 18°C, 25°C and 29°C. I used qPCR with primers flanking the intron between exons 2 and 3, called IVS3, to measure the abundance of both the spliced and unspliced *P*-element transcript relative to *rp49* housekeeping gene in the dissected ovaries of F1 offspring of these crosses (Fig. 2.3). As expected, levels of unspliced transcript are much higher than that of the spliced transcript, as splicing is mostly suppressed, even in the germline (Kofler *et al.*, 2015; Teixeira *et al.*, 2017). Regardless of the level of dysgenesis, we found that the overall expression of the *P*-

element is ~2 fold higher in a dysgenic cross (Q female x P male) compared to a genetically identical non-dysgenic cross (P female x Q male), as is also true for *D. melanogaster* (Moon *et al.*, 2018). The fold change expression of the spliced, transposase-coding version of the *P*-element is higher in the dysgenic cross, but there appear to be no gross differences in *P*-element expression in high and low tolerance lines – in the F1 offspring of most and least P tolerant lines expression of the *P*-element transcripts changes to the same degree.

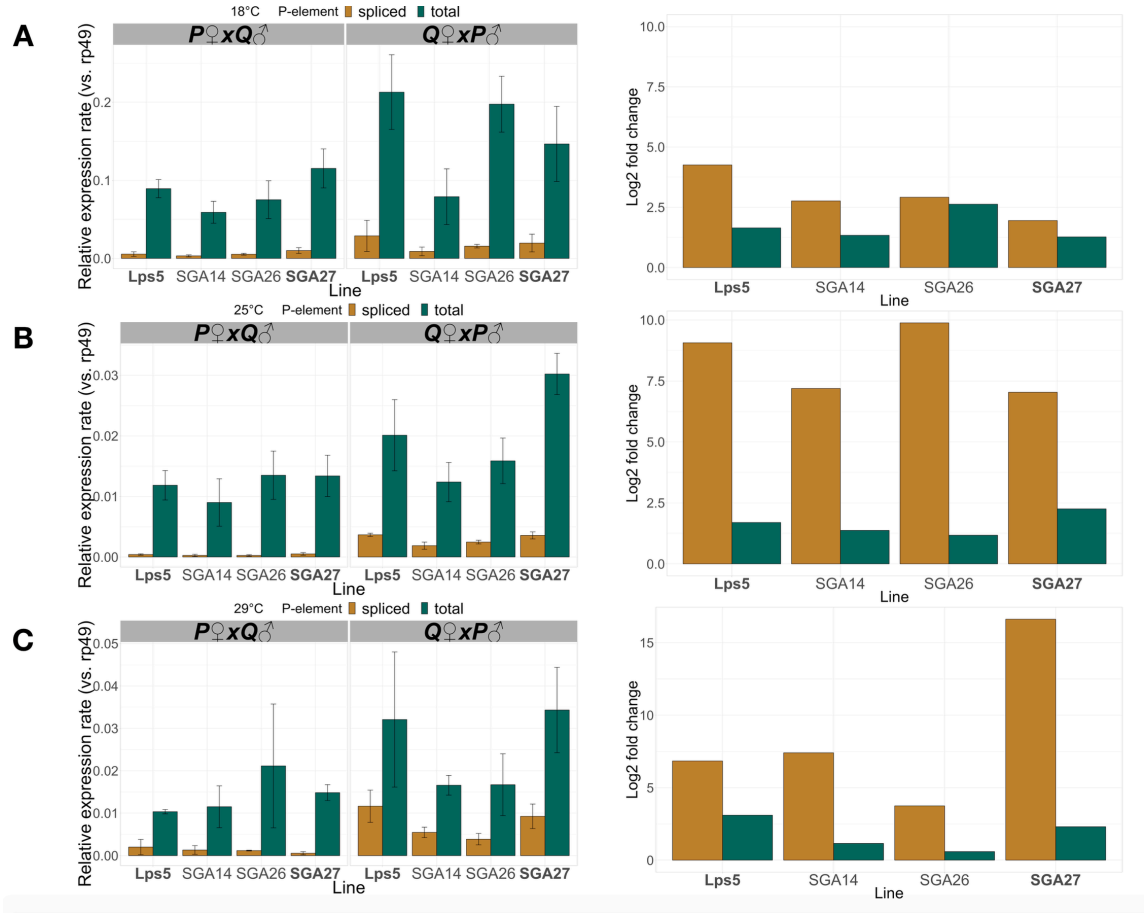


Figure 2.3. Expression rates (mean ± sd) of spliced (ochre) and total (green) *P*-element transcripts relative to *rp49* housekeeping gene at different temperatures (left) **A** 18°C **B** 25°C **C** 29°C and fold change expression of spliced (ochre) and total (green) *P*-element transcripts at each of the temperaturas between reciprocal crosses (P male x Q female and Q male x P female) (right). Isoline on X axis represents the Q line in each cross. Most tolerant lines are marked in bold.

2.3.4 Additional potential causes of variation in *P*-element tolerance

As expression of piRNAs and the regulation of splicing of the *P*-element does not explain the variation in the P tolerance, I next examined other possible causes of variation in hybrid dysgenesis between lines. One potential cause is repressive forms of *P*-elements. These are typically defective or truncated versions of *P*-elements, which can interfere with the normal activity of *P*-element transposase. The best studied of these is the KP element, which is a version of the *P*-element with an internal deletion causing a frame shift mutation (Black *et al.*, 1987). I do not expect that KP elements are likely to cause variation in maternal suppression of the *P*-element among 12 lines here, as they do not seem to be maternally acting suppressors (Wakisaka *et al.*, 2017). Nevertheless, I checked these lines for the presence of the KP element, which encodes a truncated, repressive form of *P*-element transposase (Black *et al.*, 1987; C. C. Lee *et al.*, 1996). I PCR amplified *P*-elements using tiled PCRs from the 12 lines, and sequenced them using miSeq (Hill *et al*. in prep); no breakpoints specific for the KP element occurred in these lines.

Finally, I considered the possibility that other transposable elements may be a cause of the phenotypic variation in hybrid dysgenesis. Another dysgenesis-inducing TE, *hobo*, also causes an ovarian phenotype resembling that caused by the *P*-element (Blackman *et al.*, 1987; Yannopoulos *et al.*, 1987). As all of the lines studied here contain *hobo* elements, these are not classic *hobo* dysgenesis crosses, which involve 'Empty' lines lacking *hobo* and 'Hobo' lines, in a system analogous to that of the *P*-element M and P cytotypes. However, it is possible that there is

background variation in *hobo* dysgenesis which underlies the variation in the dysgenic phenotype seen here (Srivastav and Kelleher, 2017). I do not expect variation in the hybrid dysgenesis phenotype due to differences between male parents, as all lines were crossed to the same paternal line. If so, I would expect some relationship between *hobo* copy number and the strength of dysgenesis; however, I did not find any correlation between the P tolerance and the number of *hobo* per genome as measured by qPCR (Fig. 2.4).

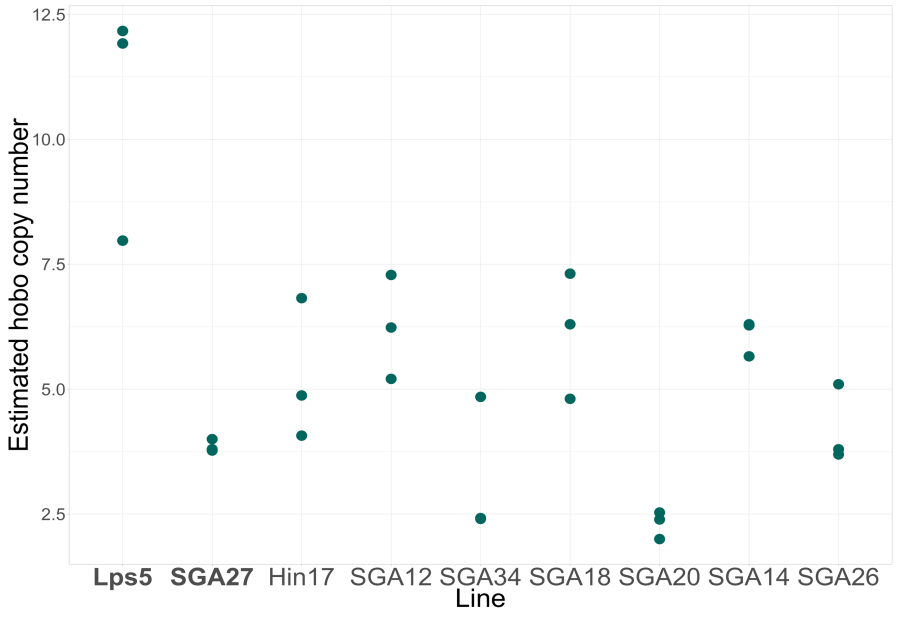


Figure 2.4. *Hobo* copy-number in some of the studied lines assessed by qPCR in some of the studied lines. Most tolerant lines are marked in bold.

2.3.5 Small RNA expression in the presence of an active *P*-element

To test the possibility that differences in piRNA expression can be seen in the presence of an active *P*-element, I selected two most tolerant – Lps5 and SGA27 – and

two least tolerant – SGA26 and SGA14 – to the *P*-element-induced hybrid dysgenesis lines. I crossed females of these lines to the same P type line as before and looked at the expression of the piRNAs in the ovaries of F1 offspring of these crosses as well as the paternal and maternal lines. I found significantly higher expression of *P*-element derived piRNAs in the P line compared to the studied Q-lines (voom/limma t=7.00, p=0.00068). However, there was no differential expression of the piRNAs between the F1 offspring from the most and least tolerant lines (voom/limma t=-3.7 p=0.88). The expression of piRNAs against the *P*-element in the presence of an active *P*-element cannot explain the variance in the degree of P tolerance in the studied lines.

I then examined whether paternal piRNAs are expressed in the ovaries of three-day old F1 daughters of P type males crossed to M type females. Previous work found severely reduced expression of piRNAs inherited from a P type father in dysgenic crosses in *D. melanogaster* (Khurana *et al.,* 2011). Similarly, I find reduced expression of *P*-element piRNAs in the ovaries of F1 daughters, compared to expression in ovaries of females of the tester P line. However, the piRNAs that were expressed by the F1 daughters did match the sequence of the paternal *P*-element piRNAs, suggesting daughters were expressing paternally inherited piRNAs clusters (Fig.2.5).

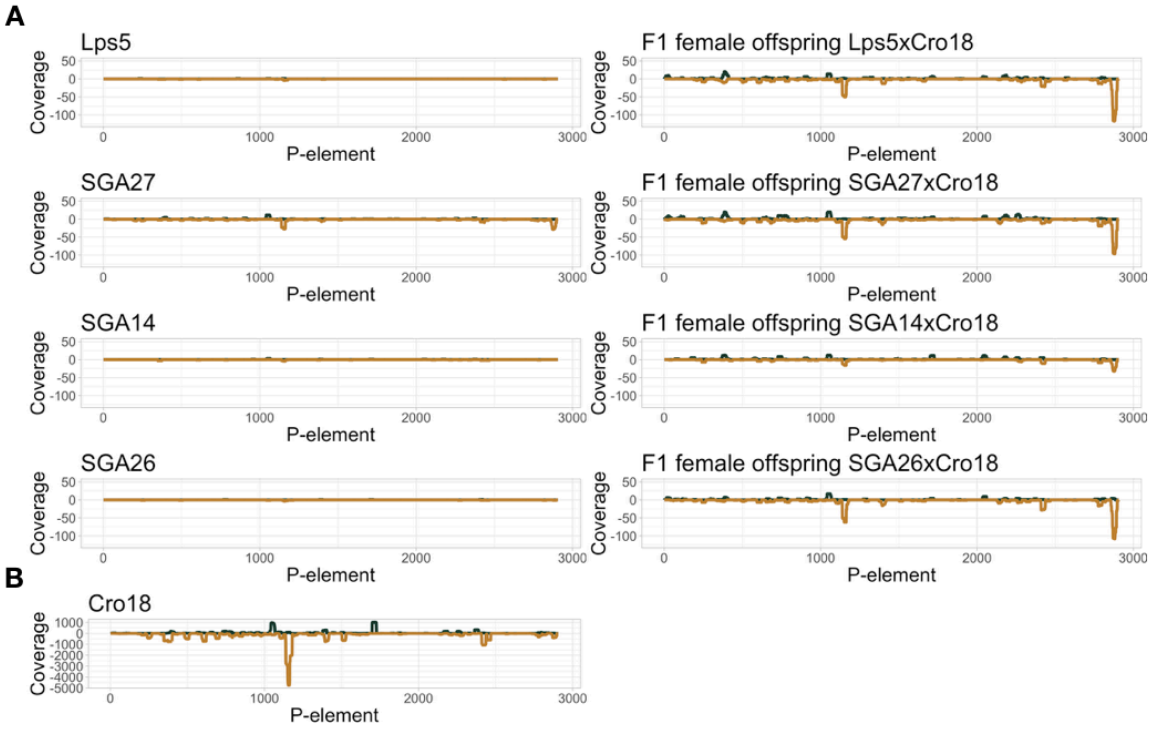


Figure 2.5. Normalised coverage of the *P*-element in small RNAseq data (**A**) four of the studied lines (left) and in the F1 females of the crosses of the lines and a tester P line (right); (**B**) in the tester P line, Cro18. Reads were filtered for min length of 24 and mapped with one mismatch allowed. Colour of the line represents the orientation (green: forward, ochre: reverse).

However, the F1 daughters do show elevated expression of *P*-element piRNAs relative to the female parents, including piRNAs that match those coming from the male parent, implying that they are expressed form the paternal clusters. Paternal piRNA expression has also been observed in *D. melanogaster*, though it is worth noting that here it occurs in 3-day old females, earlier than seen previously (Khurana *et al.*, 2011; Moon *et al.*, 2018).

2.3.6 Testing dominance of P tolerance

I crossed two most tolerant (Lps5 and SGA27) and two least tolerant lines (SGA26 and SGA14) to each other in 8 possible combinations. The F1 offspring of these crosses were crossed to tester P lines and F2 was tested for hybrid dysgenesis. The proportion of dysgenic offspring is shown in figure 6. The tolerance to HD is recessive, as the F1 offspring are similar in tolerance to their least tolerant parent (Supplementary Table 2.6). This is consistent with no effect of piRNAs on tolerance, as a piRNA driven tolerance would be expected to be dominant. Further, suppression of piRNAs shows a grandmother effect (de Vanssay *et al.*, 2012; Ronsseray, 2015). P tolerance shows no grandmother effect (FET p=0.38) as offspring of tolerant grandmothers suffer from hybrid dysgenesis to the same degree as offspring of the susceptible to hybrid dysgenesis grandmothers.

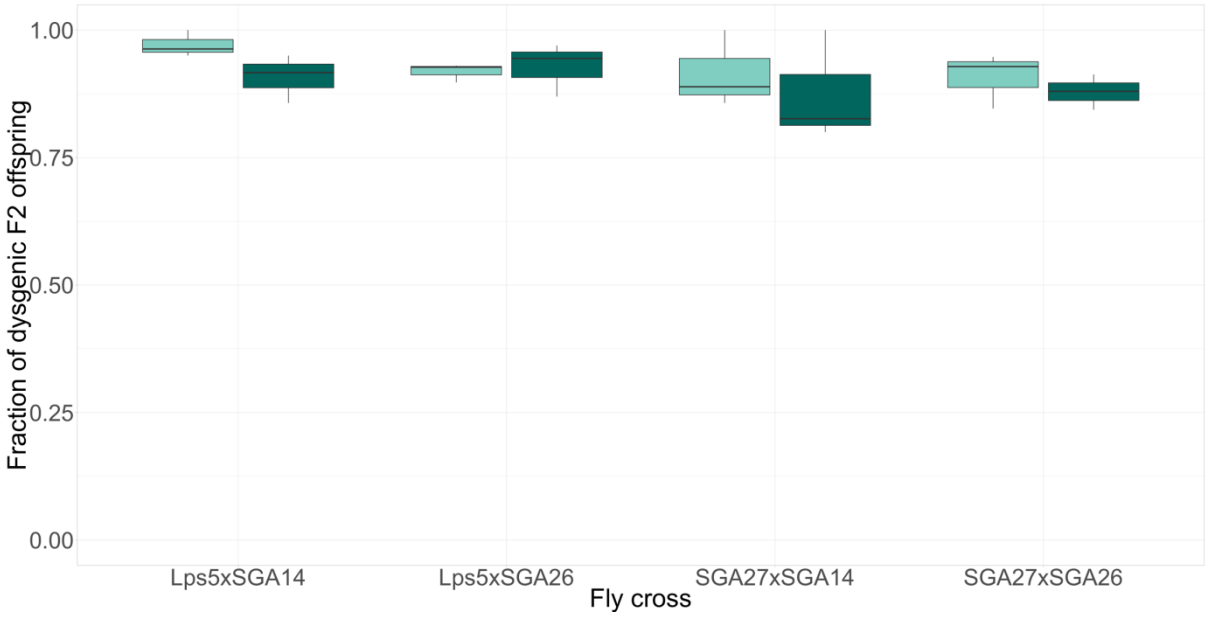


Figure 2.6. Boxplot of the proportion of hybrid dysgenesis in the F2 offspring (mean ± sd, n=30-120) when most tolerant lines (Lps5 and SGA27) were crossed to least tolerant (SGA14 and SGA26) lines (three replicates per cross). The colours indicate the direction of the cross – light green is for most tolerant grandmothers and dark green for most tolerant grandfathers.

2.4 DISCUSSION

I examined tolerance to *P*-element-induced hybrid dysgenesis *D. simulans* lines collected during the invasion of the *P*-element and observed heritable variation in tolerance. While most of the lines were vulnerable to *P*-element activity to some degree, four lines were highly tolerant, as high as in modern post-invasion P lines. Despite the variation in tolerance to the *P*-element activity, I found no evidence of differential expression of the *P*-element-derived piRNAs, or any other small RNA. Overall, my data shows no evidence that piRNAs are a major factor in the tolerance to the *P*-element in these lines caught in the early stages of *P*-element invasion of *D. simulans*.

Given the extensive evidence demonstrating that piRNAs are the most important defence against TEs in post-invasion populations, my results are surprising. In addition to many studies characterising the suppressive effect of piRNAs on transposable elements generally (reviewed in (Ozata *et al.*, 2019)), variation in levels of *P*-element-induced piRNAs is associated with the strength of suppression of hybrid dysgenesis in Q-type lines of *D. melanogaster* (Wakisaka *et al.*, 2018). In *D. simulans*, too, the evolution of *P*-element suppression in laboratory populations co-occurred with the evolution of piRNAs acting against the *P*-element (Kofler *et al.*, 2018). Consistent with this, I find abundant *P*-element piRNA expression in the P line Cro18 compared to other lines. Therefore, it is curious that the level of piRNAs expression appeared to play no role in tolerance to the *P*-element in these lines.

My data suggest not piRNAs, but alternative factors are responsible for the variation in P tolerance in the studied lines. One of the possibilities is epigenetic suppression – some lines may be more efficient in silencing TEs transcription via chromatin modifications. Both *D. simulans* and *D. melanogaster* carry epigenetic suppressors of TE expression, one of which is more highly expressed in *D. simulans* (Lee and Karpen, 2017). Other host factors that interact directly with the *P*-element (e.g., *P*-splice inhibitor (Adams *et al.*, 1997); reviewed in (Lee and Langley, 2012)) are potential sources of variation in the level of P tolerance.

Tolerance to the *P*-element-induced hybrid dysgenesis could also be due to factors that do not interact with or target the *P*-element transcript directly, but regulate the response to the damage caused by *P*-element activity. Hybrid dysgenesis is thought to be a consequence of apoptosis of developing germ line cells (Dorogova *et al.*, 2017), triggered by double-stranded breaks of DNA (DSBs) that occur during transposition.

Differences in tolerance to HD can be caused by more efficient DNA repair in the more tolerant lines. There is precedence for such variation in *Drosophila*: tolerance to UVB, which also causes double stranded DNA breaks, is positively correlated with the expression of DNA damage response genes in *D. melanogaster* (Svetec *et al.*, 2016). Alternatively, as dygenesis involves apoptosis in gonads, genes that provide general resistance to apoptosis may also provide tolerance to *P*-element damage, independent of piRNAs. Candidates for this include the genes *p53* and *bruno*, both of which explain some non-piRNA dysgenesis tolerance in *D. melanogaster* (Kelleher *et al.*, 2018; Tasnim and Kelleher, 2018). If these tolerance genes play a

general role in coevolution between hosts and their TEs, we might expect to see signatures of antagonistic coevolution at these genes, and in fact, several genes with similar roles in the regulation of female germ-line stem cells show signs of positive selection (Kelleher *et al.*, 2018). Note that signs of rapid evolution is otherwise unexpected for genes such as these with conserved functions, nor is it expected to be due to the *P*-element invasion itself, given the short time frame (Lee and Langley, 2012).

The fact that piRNAs play a major role in the later stages of TE invasion but not during the initial stages of invasion can be an example of robustness of the system. Robustness is an ability to maintain system's functionality when experiencing mutation or stimuli (Kitano, 2004). Cells have evolved several ways to achieve the same function, so that the failure of one does not lead to the failure of the whole system, so called 'diversity' or 'heterogeneity' (Kitano, 2004). The process of establishing TE defence via piRNAs can be compared to the immune response. First, in the early stages of invasion, mechanisms not specific to a given TE act to mask negative consequences of transposition and to ensure survival of the cells ('innate' response); later on, as in adaptive immune response, piRNAs are produced, they target a specific TE and serve as a 'memory' of previous TE invasions. It is possible that pre-existing variation in HD is a factor that enables a TE to invade a population, if there are no mechanisms to cope with negative consequences of transposition in the host, the invasion may be not possible due to the cost it imposes on the host. The full complement of molecular mechanisms that underlie pre-adaptation and response to newly invading TEs have yet to be fully identified.

Chapter 3

3.1 INTRODUCTION

The most widespread example of selfish genetic elements is transposable elements, genes that can change their location within a genome ('transpose') and therefore increase in copy number. Transposable elements are abundant in nearly all eukaryotic species examined by this time and can form up to ~95% of some plant genomes (Arkhipova and Meselson, 2000; Gregory, 2005; Kronmiller and Wise, 2008; de Koning *et al.*, 2011). Transposition is a cause of mutations that are sometimes beneficial, but usually deleterious: it can disrupt essential genes, lead to ectopic recombination, and double-stranded breaks that are toxic to cells (Noutsopoulos *et al.*, 2010; Dorogova *et al.*, 2017). One of the best studied examples of the deleterious consequences of transposition in *Drosophila* is hybrid dysgenesis caused by *P*-element transposition. Hybrid dysgenesis (Pasyukova *et al.*, 2004) is a phenomenon associated with high mutation rates, ectopic recombination and sterility. HD happens as a result of inability of a female lacking *P*-elements in her genome to control *P*-element activity in an egg fertilised by a male with several copies of the *P*-element. The reason is that these females are missing a class of small non-coding RNAs that target the P-element. In particular, they lack *P*-element derived PIWI-interacting RNAs, or 'piRNAs'. These piRNAs are a major factor protecting the germ line of animals against the negative consequences of transposition. They are loaded into the egg by the female parent to protect their offspring (Brennecke *et al.*, 2007). piRNAs originate from a number of mostly heterochromatic regions of the genome, piRNA clusters (Brennecke *et al.*, 2007).

piRNA clusters consist of degenerate copies of TEs, and piRNAs work by targeting TE transcripts based on sequence complementarity. Thus, individual flies only produce piRNAs against the TEs that have homology to those that reside in their piRNA clusters.

Transposable elements not only relocate within one genome, but also invade new species. *P*-element is a best example of this occurring – it invaded and spread in in *Drosophila melanogaster* in the 20th, and in the sister species *D. simulans* in the 21st century. Interestingly, as we saw in the previous Chapter, some *D. simulans* lines collected during the process of invasion of the *P*-element tolerate *P*-element transposition despite not producing *P*-element derived piRNAs and not having *P*-elements in their genomes (Hill *et al.*, 2016; Chapter 2). In *D. melanogaster*, tolerance to uncontrolled *P*-element activity was mapped to the *bruno* locus, a developmental regulator of oogenesis (Kelleher *et al.*, 2018). This QTL explains ~35% of the heritable variation in tolerance to the P-element, which suggests there might be other genetic loci and factors involved in the ability to tolerate consequences of uncontrolled TE transposition.

Here, I examined two most and two least tolerant to HD lines of *D. simulans*, all of which lack piRNAs targeting the *P*-element. To further investigate this issue and look at the genetic basis of the tolerance, I performed a set of crosses between the most and least tolerant to HD lines of *D. simulans* (F1, F2, maternal and paternal backcrosses). I estimated how many loci underlie the phenotypic differences observed, and ask whether these loci are dominant or recessive. The results suggest that loci involved in the tolerance are unlikely to be fully dominant or fully recessive.

3.2 MATERIALS AND METHODS

3.2.1 Fly stocks

The *D. simulans* isofemale lines used in this study were collected in 2009 in Athens and Morven, Georgia, USA (by P. Haddrill and A. Paaby). They differ in their phenotype – tolerance to the *P*-element induced hybrid dysgenesis, with two of the lines (Lps5 and SGA27) being highly tolerant to *P*-element induced HD, and the other two (SGA14 and SGA26) highly susceptible to HD (Chapter 2). Flies were maintained on cornmeal-molasses-yeast-agar *Drosophila* medium at 25 °C.

3.2.2 Crosses

For each of the crosses, I crossed 5 virgin females to at least 2 males. I crossed each of the most tolerant to HD lines to each of the least tolerant to HD lines in both directions and two replicates. I then crossed F1 offspring to each other and obtained F2. I also performed maternal and paternal backcrosses, then male F1 offspring was backcrossed to the paternal line and female F1 offspring to maternal line, respectively. The crossing scheme is shown in Figure 3.1.

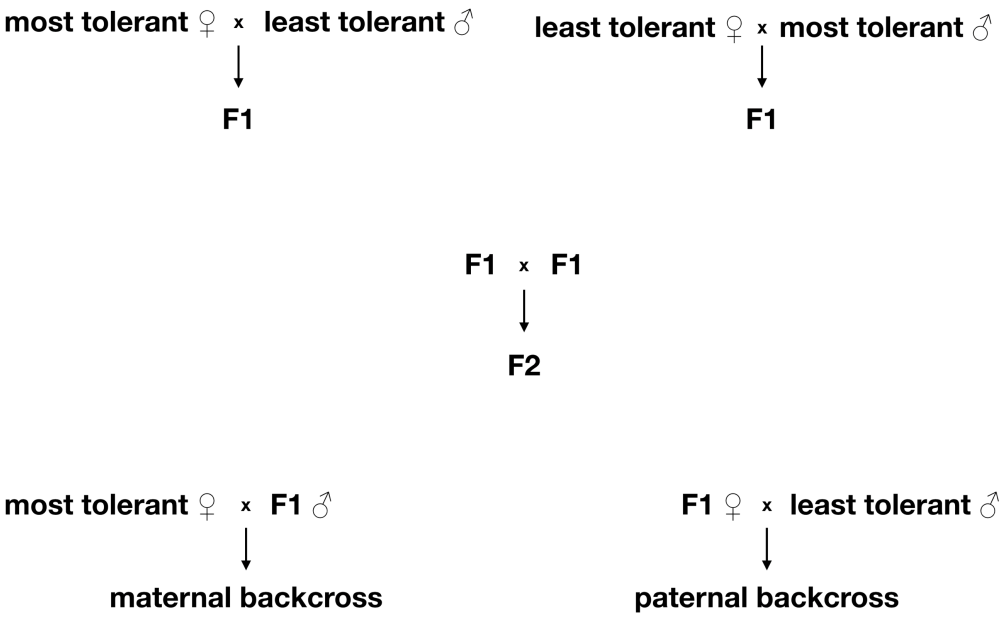


Figure 3.1. Scheme representing the crosses set up to estimate the number of loci underlying tolerance to HD.

3.2.3 Assaying hybrid dysgenesis

I performed the crosses at 29 °C, the temperature at which HD can be seen. From each of the crosses performed, I took female offspring (F1, F2, maternal and paternal backcrosses) and crossed it to a *D. simulans* line known to induce hybrid dysgenesis — Cro18 (Chapter 2). All of the female offspring of these crosses were tested for HD — only females with two well-developed ovaries were considered to be normal and therefore tolerant, the rest was counted as 'dysgenic'.

3.2.4 Calculating expectations under different genetic models

I considered seven possible scenarios for the number of the loci involved in the genetic basis of tolerance (Table 3.1). The maximum number of loci in these models correspond to the maximum number of major linkage groups in *D. simulans.* For each of these scenarios, I examine different dominances of the loci involved, from completely recessive (dominance = 0) to completely dominant (dominance = 1). For the cases of one locus (whether autosomal or X-linked), we look at dominances at 0.1 step (0, 0.1, 0.2, …), for all other cases at 0.25 step (0, 0.25, 0.5, …). I used a custom python script that calculates the expected proportions of tolerant phenotypes based on a number of loci involved and their dominance. I only consider cases where all of the loci involved have the same effect (1/total number of loci considered).

	Number of loci	
Model	**Autosomal**	**X-linked**
1	1	
2	2	
3	3	
4	4	
5		1
6	1	1
7	2	1

Table 3.1. Seven scenarios for the number of loci involved in tolerance to *P*-element activity.

3.2.5 Data analysis

First, I checked to see whether different tolerant and susceptible lines behaved differently. To do this, I analysed the data to ask whether there was a significant 'line' effect. Analysis with a binomial GLM indicated that the data were overdispersed, so I used a quasi-binomial GLM that allows for additional variance in the data. I then checked whether the number of tolerant offspring can be predicted by the cross type (most tolerant ♀ x least tolerant ♂ or least tolerant ♀ x most tolerant ♂), line (SGA14, SGA26, SGA27 or Lps5) and generation (F1, F2, BC maternal and paternal). No significant effect was found ($p = 0.26$).

To compare the observed data to the expected data for each model, I used the chi-squared deviation to calculate the differences between expected and observed values. For each of the number of loci and dominance combinations (e.g. one locus of dominance 0, one locus of dominance 0.1, etc.), I calculated the sum of the chi-squares over all crosses and generations.

3.3 Results and Discussion

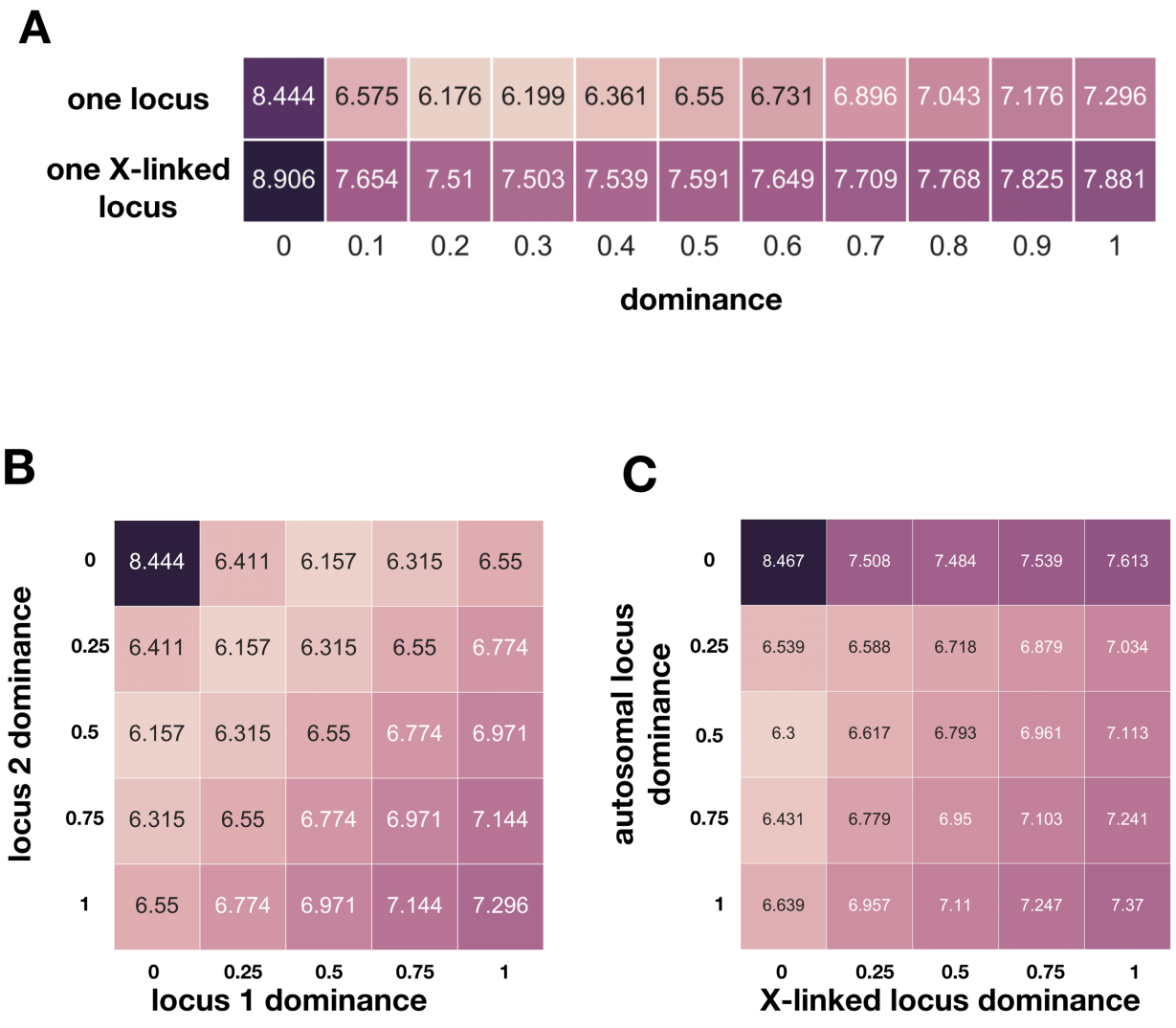
I observed heritable variation in the tolerance to *P*-element induced hybrid dysgenesis that cannot be explained by the expression of piRNAs or any other kind of small RNAs (Chapter 2). I set up crosses to estimate the genetic basis of the tolerance (Fig. 3.1). I crossed each of the most tolerant to HD isolines (Lps5, SGA27, showing 80-90% of normal offspring in dysgenic direction of the cross) to each of the least tolerant to HD isolines (SGA14, SGA26, showing ~10% of normal offspring in the dysgenic cross) (Fig. 3.1). I tested female offspring of each of the generations (F1, F2, maternal and paternal backcrosses) for tolerance to HD by crossing them to the HD-inducing line of *D. simulans*, Cro18.

I calculated the number of expected tolerant offspring in each of the crosses based on the number and dominance of the loci involved in the genetics of tolerance. I considered several possibilities for the number of loci, either X-linked or autosomal (Table 3.1). I assumed Mendelian inheritance of each of the loci, with all loci unlinked. In cases where more than one locus is considered, all loci had equal effects, equal to 1/number of loci. For each case, each locus has a dominance in the range from 0 (completely recessive) to 1 (fully dominant). For my predictions, I do not take into account penetrance of the trait or epistasis.

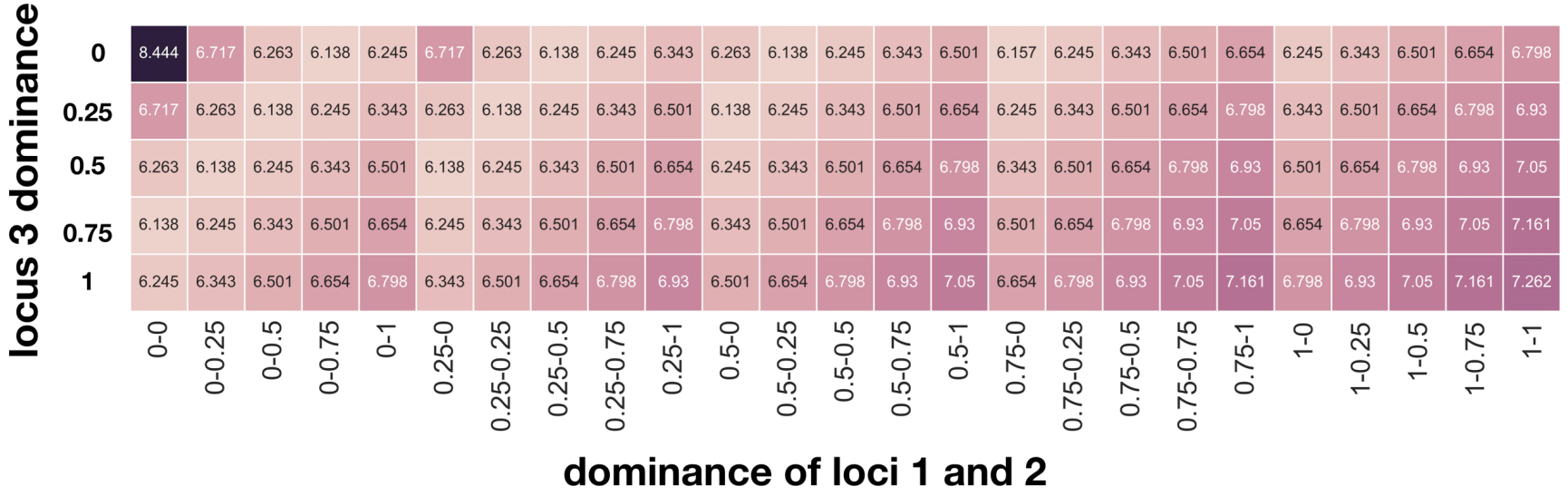
The results showed that cross type (F1, F2, maternal or paternal backcrosses) or line cannot predict the number of tolerant offspring (p=0.26). I use the chi-square deviation to calculate the differences between numbers of observed and expected values of tolerant offspring. Specifically, for each of the cases of number of loci and dominance combinations, I summed chi-square values for all of the types and

generations of crosses. I used this chi-square deviation to see which models best fit the outcomes of the crosses (Fig. 3.2).

Figure 3.2. Heatmaps showing the log of chi-square sums between the observed and predicted values for each of the cases considered. Numbers in the squares indicate the logs of the chi-square sums over all crosses and generations. The log of chi-square sums shown in squares were calculated as follows: I calculated the differences between the numbers of observed and expected dysgenic offspring for each of the cases of dominance and number of loci involved. Next, I summed chi-square values for all of the types and generations of crosses and took log of the obtained values. **A** One autosomal locus (top row) and one X-linked locus (bottom row) **B** Two autosomal loci of equal 0.5 effect **C** One autosomal and one X linked locus, each having 0.5 effect **D** Three autosomal loci, each of the effect 0.33 **E** Two autosomal and one X-linked locus **F** Four autosomal loci, effect of each of them 0.25. Lower logs of chi-square sums indicate better fit of the data to the model.

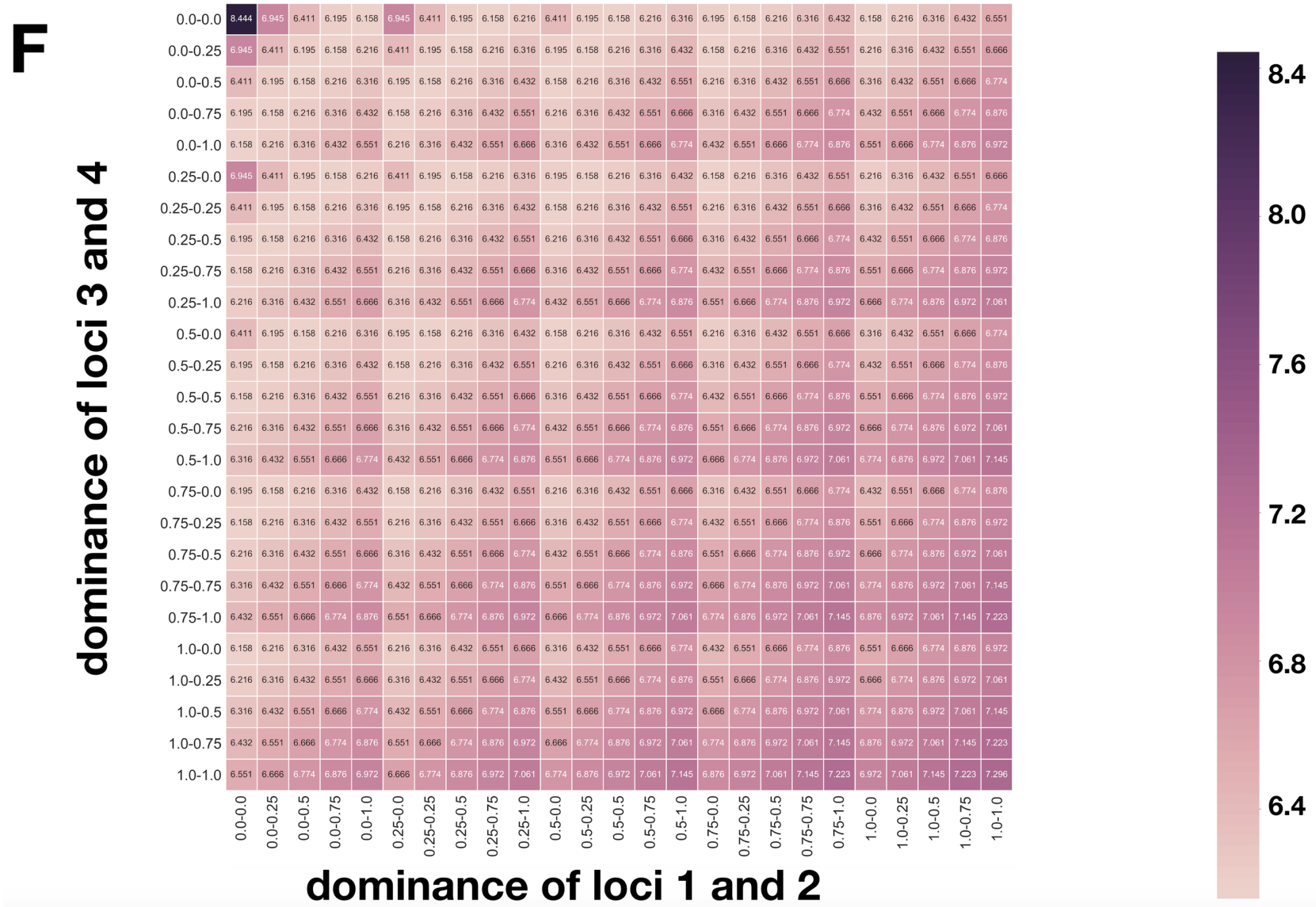


D


locus 3 dominance
0
0.25
0.5
0.75
1
8.444 6.717 6.263 6.138 6.245 6.717 6.263 6.138 6.245 6.343 6.263 6.138 6.245 6.343 6.501 6.157 6.245 6.343 6.501 6.654 6.245 6.343 6.501 6.654 6.798
6.717 6.263 6.138 6.245 6.343 6.263 6.138 6.245 6.343 6.501 6.138 6.245 6.343 6.501 6.654 6.245 6.343 6.501 6.654 6.798 6.343 6.501 6.654 6.798 6.93
6.263 6.138 6.245 6.343 6.501 6.138 6.245 6.343 6.501 6.654 6.245 6.343 6.501 6.654 6.798 6.343 6.501 6.654 6.798 6.93 6.501 6.654 6.798 6.93 7.05
6.138 6.245 6.343 6.501 6.654 6.245 6.343 6.501 6.654 6.798 6.343 6.501 6.654 6.798 6.93 6.501 6.654 6.798 6.93 7.05 6.654 6.798 6.93 7.05 7.161
6.245 6.343 6.501 6.654 6.798 6.343 6.501 6.654 6.798 6.93 6.501 6.654 6.798 6.93 7.05 6.654 6.798 6.93 7.05 7.161 6.798 6.93 7.05 7.161 7.262
0-0
0-0.25
0-0.5
0-0.75
0-1
0.25-0
0.25-0.25
0.25-0.5
0.25-0.75
0.25-1
0.5-0
0.5-0.25
0.5-0.5
0.5-0.75
0.5-1
0.75-0
0.75-0.25
0.75-0.5
0.75-0.75
0.75-1
1-0
1-0.25
1-0.5
1-0.75
1-1
dominance of loci 1 and 2

X-linked locus dominance
dominance of loci 1 and 2
0.0
0.25
0.5
0.75
1.0
0.0-0.0
0.0-0.25
0.0-0.5
0.0-0.75
0.0-1.0
0.25-0.0
0.25-0.25
0.25-0.5
0.25-0.75
0.25-1.0
0.5-0.0
0.5-0.25
0.5-0.5
0.5-0.75
0.5-1.0
0.75-0.0
0.75-0.25
0.75-0.5
0.75-0.75
0.75-1.0
1.0-0.0
1.0-0.25
1.0-0.5
1.0-0.75
1.0-1.0
8.456 6.778 6.329 6.247 6.326 6.778 6.329 6.247 6.326 6.46 6.329 6.247 6.326 6.46 6.608 6.247 6.326 6.46 6.608 6.752 6.326 6.46 6.608 6.752 6.886
7.523 6.526 6.319 6.347 6.46 6.526 6.319 6.347 6.46 6.599 6.319 6.347 6.46 6.599 6.74 6.347 6.46 6.599 6.74 6.873 6.46 6.599 6.74 6.873 6.997
7.412 6.518 6.412 6.479 6.599 6.518 6.412 6.479 6.599 6.732 6.412 6.479 6.599 6.732 6.862 6.479 6.599 6.732 6.862 6.985 6.599 6.732 6.862 6.985 7.099
7.409 6.594 6.536 6.615 6.732 6.594 6.536 6.615 6.732 6.856 6.536 6.615 6.732 6.856 6.975 6.615 6.732 6.856 6.975 7.088 6.732 6.856 6.975 7.088 7.193
7.441 6.699 6.665 6.746 6.856 6.699 6.665 6.746 6.856 6.97 6.665 6.746 6.856 6.97 7.08 6.746 6.856 6.97 7.08 7.183 6.856 6.97 7.08 7.183 7.281

F


dominance of loci 3 and 4
0.0-0.0 8.444 6.945 6.411 6.195 6.158 6.945 6.411 6.195 6.158 6.216 6.411 6.195 6.158 6.216 6.316 6.195 6.158 6.216 6.316 6.432 6.158 6.216 6.316 6.432 6.551
0.0-0.25 6.945 6.411 6.195 6.158 6.216 6.411 6.195 6.158 6.216 6.316 6.195 6.158 6.216 6.316 6.432 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666
0.0-0.5 6.411 6.195 6.158 6.216 6.316 6.195 6.158 6.216 6.316 6.432 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774
0.0-0.75 6.195 6.158 6.216 6.316 6.432 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876
0.0-1.0 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972
0.25-0.0 6.945 6.411 6.195 6.158 6.216 6.411 6.195 6.158 6.216 6.316 6.195 6.158 6.216 6.316 6.432 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666
0.25-0.25 6.411 6.195 6.158 6.216 6.316 6.195 6.158 6.216 6.316 6.432 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774
0.25-0.5 6.195 6.158 6.216 6.316 6.432 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876
0.25-0.75 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972
0.25-1.0 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972 6.666 6.774 6.876 6.972 7.061
0.5-0.0 6.411 6.195 6.158 6.216 6.316 6.195 6.158 6.216 6.316 6.432 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774
0.5-0.25 6.195 6.158 6.216 6.316 6.432 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876
0.5-0.5 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972
0.5-0.75 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972 6.666 6.774 6.876 6.972 7.061
0.5-1.0 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972 6.666 6.774 6.876 6.972 7.061 6.774 6.876 6.972 7.061 7.145
0.75-0.0 6.195 6.158 6.216 6.316 6.432 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876
0.75-0.25 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972
0.75-0.5 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972 6.666 6.774 6.876 6.972 7.061
0.75-0.75 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972 6.666 6.774 6.876 6.972 7.061 6.774 6.876 6.972 7.061 7.145
0.75-1.0 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972 6.666 6.774 6.876 6.972 7.061 6.774 6.876 6.972 7.061 7.145 6.876 6.972 7.061 7.145 7.223
1.0-0.0 6.158 6.216 6.316 6.432 6.551 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972
1.0-0.25 6.216 6.316 6.432 6.551 6.666 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972 6.666 6.774 6.876 6.972 7.061
1.0-0.5 6.316 6.432 6.551 6.666 6.774 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972 6.666 6.774 6.876 6.972 7.061 6.774 6.876 6.972 7.061 7.145
1.0-0.75 6.432 6.551 6.666 6.774 6.876 6.551 6.666 6.774 6.876 6.972 6.666 6.774 6.876 6.972 7.061 6.774 6.876 6.972 7.061 7.145 6.876 6.972 7.061 7.145 7.223
1.0-1.0 6.551 6.666 6.774 6.876 6.972 6.666 6.774 6.876 6.972 7.061 6.774 6.876 6.972 7.061 7.145 6.876 6.972 7.061 7.145 7.223 6.972 7.061 7.145 7.223 7.296
0.0-0.0 0.0-0.25 0.0-0.5 0.0-0.75 0.0-1.0 0.25-0.0 0.25-0.25 0.25-0.5 0.25-0.75 0.25-1.0 0.5-0.0 0.5-0.25 0.5-0.5 0.5-0.75 0.5-1.0 0.75-0.0 0.75-0.25 0.75-0.5 0.75-0.75 0.75-1.0 1.0-0.0 1.0-0.25 1.0-0.5 1.0-0.75 1.0-1.0
dominance of loci 1 and 2
8.4
8.0
7.6
7.2
6.8
6.4

My results likely exclude several possibilities of the genetic basis of the tolerance to HD observed between most and least tolerant lines of *D. simulans*. Tolerance does not seem to be a fully dominant trait (which is in agreement with previous findings, chapter 2). It also does not appear to be a fully recessive trait with dominance of 0 independent of the number of loci considered.

I appear to have little power overall to discriminate between models with different numbers of genetic loci. However, models with more loci do fit the observed data more closely than those with fewer loci. It is likely *a priori* that more than one locus underlies the genetic basis of tolerance. In *D. melanogaster*, a locus linked to tolerance to the *P*-element can explain ~35% of the variation in tolerance to HD (Kelleher *et al.*, 2018). I note that models that include an X-linked locus seem to be a slightly worse fit than the corresponding autosomal only models, particularly in the case of a single X-linked locus. Based on my results, I can speculate that genes responsible for the phenotype do not seem to be X-linked, whether I consider cases of one X-linked locus or several — one X-linked and other autosomal. If I can exclude an entire chromosome as contributing to the phenotype, then this suggests that the trait, though not due to a single locus, is also not highly polygenic.

In conclusion, my results allow me hypothesise that tolerance to the *P*-element is moderately polygenic, and that alleles at the genes involved in the trait do not seem to be fully recessive or fully dominant. However, I cannot tell exactly how many loci underlie the tolerance to HD for several reasons. I calculated my expected numbers assuming that most and least tolerant lines showed 100% and 0% of

normal offspring, respectively; however, this is not quite true in reality as none of the lines are fully tolerant or not. Also, for my calculations I made a range of assumptions: unlinked loci of incomplete penetrance, loci involved in it can be in linkage disequilibrium and there might be epistasis, negative or positive, between them.

To get a better estimation of the genetic basis of tolerance, I could perform several generations of inbreeding prior to performing crosses to homogenous the parental lines. It would also be preferable to use marked lines, so that the origins of each chromosomes can be tracked. There are *D. simulans* lines with visible markers that could be used. However, to increase the resolution, standard QTL mapping techniques should be used. Specifically, an advanced intercross mapping method designed specifically for *Drosophila*. This would involve crossing of three most tolerant lines to a single least tolerant line, splitting the offspring into 10 sublines and intercrossing these for 15 generations. Next, females from the final cross should be assayed for P tolerance and the most tolerant ones sequenced alongside the parental lines. This approach should have good power to identify causative loci.

In conclusion, previous work found that *D. simulans* lines collected during the early stages of the *P*-element invasion show heritable variation in tolerance to *P*-element induced hybrid dysgenesis. Results from the previous chapter have shown that P tolerance cannot be explained by piRNA expression. Here, I show tentative evidence that P tolerance is a polygenic trait, likely to be autosomal.

Chapter 4

4.1 INTRODUCTION

Selfish genetic elements are elements that can invade, spread and persist in populations despite not contributing to their host organism's fitness, or even despite being harmful to the host (Orgel and Crick, 1980; Arkhipova and Meselson, 2000; Burt and Trivers, 2006; Wicker *et al.*, 2007). Transposable elements are the most widespread examples of selfish elements, and they have invaded and spread in essentially all of the eukaryotic genomes examined so far (Orgel and Crick, 1980; Arkhipova and Meselson, 2000; Burt and Trivers, 2006; Wicker *et al.*, 2007). TEs can become extremely common in their eukaryotic hosts' genomes, accounting for up to 95% of some plant genomes, and so can be major determinants of host genome size (Gregory, 2005; Kronmiller and Wise, 2008; Canapa *et al.*, 2015). A key feature of transposable element is the ability to transpose – to change their location within genomes and produce copies of themselves. Transposition of TEs into new locations is a source of mutations, which are often deleterious, and it can result in the disruption of essential genes and chromosome breakage (Burt and Trivers, 2006). In addition, TE insertions can have other harmful effects, such as structural rearrangement and ectopic recombination (Langley *et al.*, 1988; Hua-Van *et al.*, 2011; Chuong *et al.*, 2017). These deleterious consequences of transposition have forced animals to evolve a variety of mechanisms to protect their genomes, such as DNA and chromatin modifications, and some forms of RNA interference (Slotkin and Martienssen, 2007).

One of the most important controls of TE activity, which silences transposable elements in the germ line, is due to PIWI-interacting RNAs (piRNAs) (Brennecke *et al.*, 2007; Aravin *et al.*, 2008; Khurana *et al.*, 2011; Iwasaki *et al.*, 2015). piRNAs are small non-coding RNAs, normally between 24 and 36 nt long, which are expressed mainly in the gem line (Vagin *et al.*, 2006; Aravin *et al.*, 2007; Brennecke *et al.*, 2007; Houwing *et al.*, 2007; Kuramochi-Miyagawa *et al.*, 2008). piRNA do not have any particular sequence motif, except for a bias for uracil at their 5' end (Brennecke *et al.*, 2007). Several genomic regions, usually in the heterochromatin, serve as a source of piRNAs; these regions consist of degraded transposable elements copies and are called 'piRNA clusters' (Brennecke *et al.*, 2007). Importantly, a single insertion of a transposable element into one of the piRNA clusters leads to piRNA production and may be sufficient to silence this TE (Ronsseray *et al.*, 1991; Josse *et al.*, 2007; Zanni *et al.*, 2013).

piRNA clusters are present in many species, including *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus* and *Homo sapiens* (Aravin *et al.*, 2007; Yamanaka *et al.*, 2014; Czech and Hannon, 2016; Lewis *et al.*, 2018). In *D. melanogaster*, where piRNA clusters were initially discovered, piRNA clusters account for at least 3.5% of the genome (Brennecke *et al.*, 2007). They vary in size and can be several hundred kilobases long. In *D. melanogaster*, there are two types: uni-strand clusters that are transcribed from one genomic strand and are expressed in somatic support cells surrounding the germ line, and dual-strand clusters that are transcribed from both genomic strands and expressed in the oocyte (Brennecke *et al.*, 2007; Huang *et al.*, 2017). Both types of clusters are epigenetically marked with

H3K9me3 (histone 3 lysine 9 tri-methylation). This mark usually transcriptionally silences regions of the genome, and is commonly found in heterochromatic regions of the genome (Le Thomas *et al.*, 2013; Klenov *et al.*, 2014; Mohn *et al.*, 2014; Z. Zhang *et al.*, 2014). Uni-strand piRNA clusters are thought be be transcribed as long mRNA precursors that are processed into 24-36-nt piRNAs (Le Thomas *et al.*, 2013). These piRNAs are bound by members of the PIWI protein family, forming ribonuclearprotein complexes that target and degrade transposable element mRNA (Brennecke *et al.*, 2007). As a result of this and other piRNA mediated silencing mechanisms (Ozata *et al.*, 2019), the target TEs are expected to have highly reduced transposition rates.

To date, piRNA clusters have been described in *D. melanogaster* (Brennecke *et al.*, 2007), but, with the exception of the *flamenco* cluster, not in other *Drosophila* species. Identifying piRNA clusters in other *Drosophila* species may provide insights into piRNA cluster evolution. Here, I identify piRNA clusters in two strains of *D. simulans*, a sister species of *D. melanogaster*. To this end, I use long-read genome sequencing to assemble their genomes, including most of the problematic repetitive regions. I also use small RNA sequencing data from the same strains to identify the piRNA clusters in these assembled genomes.

4.2 Materials and Methods

4.2.1 Strains used

The *D. simulans* isofemale lines used in this study were collected in 2009 in Athens and Morven, Georgia, USA (by P. Haddrill and A. Paaby). They differ substantially in their phenotype – tolerance to the *P*-element induced hybrid dysgenesis, with one of the lines (Lps5) being highly tolerant to *P*-element induced HD, and the other one (SGA26) highly susceptible to HD (Chapter 2). Flies were maintained on cornmeal-molasses-yeast-agar *Drosophila* medium at 25 °C.

4.2.2 Inbreeding and PCR

For inbreeding, I set up ten one pair brother-sister crosses for each of the lines. After three generations of inbreeding, I extracted DNA from 10 flies of each subline using DNeasy Blood & Tissue Kit by Qiagen. I PCR amplified and sequenced 10 genes across the genome (2 for each chromosome arm) for each of the 20 sublines; primers and PCR conditions are listed in supplementary table 4.2. From each of the lines, I selected one subline that was homozygous for all sequenced genes for PacBio sequencing.

4.2.3 DNA extraction

For each subline selected for sequencing, I extracted high molecular weight (HMW) DNA from 50 female pupae. I chose to sequence only female individuals to have the same depth of sequencing for autosomes and X chromosome; I decided to

use pupae for sequencing since pupae do not have gut symbionts and therefore, this would reduce the number of reads of non-fly origin.

To extract DNA, I ground pupae in 500ul of extraction buffer (50 mM Tris pH 8.0, 25 mM NaCl, 25 mM EDTA, 0.1% SDS) with 10ul of Proteinase K (20 mg/ml) and 10 ul of RNase A; and incubated the samples for 4h at 60 °C. Next, I added 500ul of phenol to each of the samples and centrifuged them at 14 000 rpm at room temperature (RT) for 10 min. After centrifugation, I mixed 360 µl of the upper phase with 300 µl of chloroform and again centrifuged as previously. I transferred ~340 µl of the upper phase to a fresh tube with 850 µl pure ethanol and 30 µl 3M NaoAc pH 5.2; and left the mixture for DNA precipitation overnight at -20 °C. The following day I centrifuged the samples at 14 000 rpm at 4 °C for 30 min, washed the pellet twice with 70% ethanol and diluted in 110 ul Tris-EDTA pH 8.0 buffer. I checked the quality of the extracted DNA by running the samples on 0.5% agarose gel overnight at 30V with Quick load 1kb Extend DNA ladder (New England Biolabs). After quality control, I submitted DNA samples to the Centre for Genomic Research, University of Liverpool facility for library preparation and sequencing on the Sequel System.

4.2.4 Genome assembly

I assembled the obtained sequencing reads using *canu* (Koren *et al.* 2017) with settings for genome size 180 Mb and minimum read length 200 nt. I generate basic assembly statistics with BBMap (Bushnell 2015), and aligned the assembled contigs to the reference genome of *D. simulans* using *nucmer* (Marcais *et al.*, 2018). I merged contigs that aligned to the same chromosome into one scaffold with 100 Ns

between the contigs. I visualised the alignments using *assemblytics* (Nattestad and Schatz, 2016). I prepared reference genomes for each of the strains by merging chromosome scaffolds and unaligned contigs into a new reference fasta file.

4.2.5 piRNA mapping

First, I mapped small RNA reads obtained from previous experiments (chapter 2) to *D. simulans* genes from Flybase (ftp://ftp.flybase.net, dsim_r2.02_FB2017_04) and removed all the reads that mapped to the sense strand of the annotated genes, as these may be degraded mRNAs. I used the remaining reads for each of the lines and mapped them to each of the genomes using *bwa aln* v 0.7.13 (Li and Durbin, 2009) allowing for no mismatches. After mapping, I removed the reads with indels and shorter than 23 nt, which allowed filtering of reads specifically for piRNAs. I restricted our analysis to only the ~ 18% of the piRNAs that mapped uniquely to the assembled genomes, following the approach described in Brennecke *et al.* 2007.

4.2.6 TE density in the assembled genomes

I used Repeatmasker (Smit *et al.* 2017) to mask *Drosophila* transposable elements (annotation v. 9.42 available from http://flybase.org/) in the assembled genomes. I then calculated density of the TEs in the genomes as a percentage of TE bases in 50 kb non-sliding windows.

4.2.7 piRNA cluster identification

piRNA clusters consist of nested transposable elements and therefore, most piRNAs will match several locations within a cluster and the genome. Therefore, to identify the locations of piRNA clusters, I restricted our analysis to ~18% of the piRNAs that map uniquely, following Brennecke *et al.* (2007). Note that the same strains were used for both small RNA and PacBIO sequencing here, increasing the chance that the location of a uniquely mapped piRNA reflect the true origin of a small RNA. I traversed the genome in 5kb sliding windows, and identified all of the regions that had more than 1 piRNA per window. I used a cut-off of 1000 uniquely mapping piRNAs per window to identify the regions of potential piRNA clusters. Clusters located within 20 kb from each other were merged.

4.2.8 piRNA cluster comparison with *D. melanogaster*

I compared two of the major clusters in *D. melanogaster* (*flamenco* and *42AB*) to the corresponding clusters in *D. simulans*. *Flamenco* in *D. melanogaster* is located downstream the *DIP-1* gene, and we used this gene to locate *flamenco* in the *D. simulans* genome. For *42AB*, the location was determined by the position of the cluster in *D. melanogaster* and corresponding position in *D. simulans* genome. I used custom python scripts to extract the sequences of the clusters from genome files. The scripts are located at https://github.com/OlgaPawlowska/python-scripts. After that, I used *Repeatmasker* (Smit *et al.* 2017) to identify TE sequences within each

cluster for + and - strands. I then assigned each TE family to a class (Wicker *et al.* 2007) (Diagram 4.1).

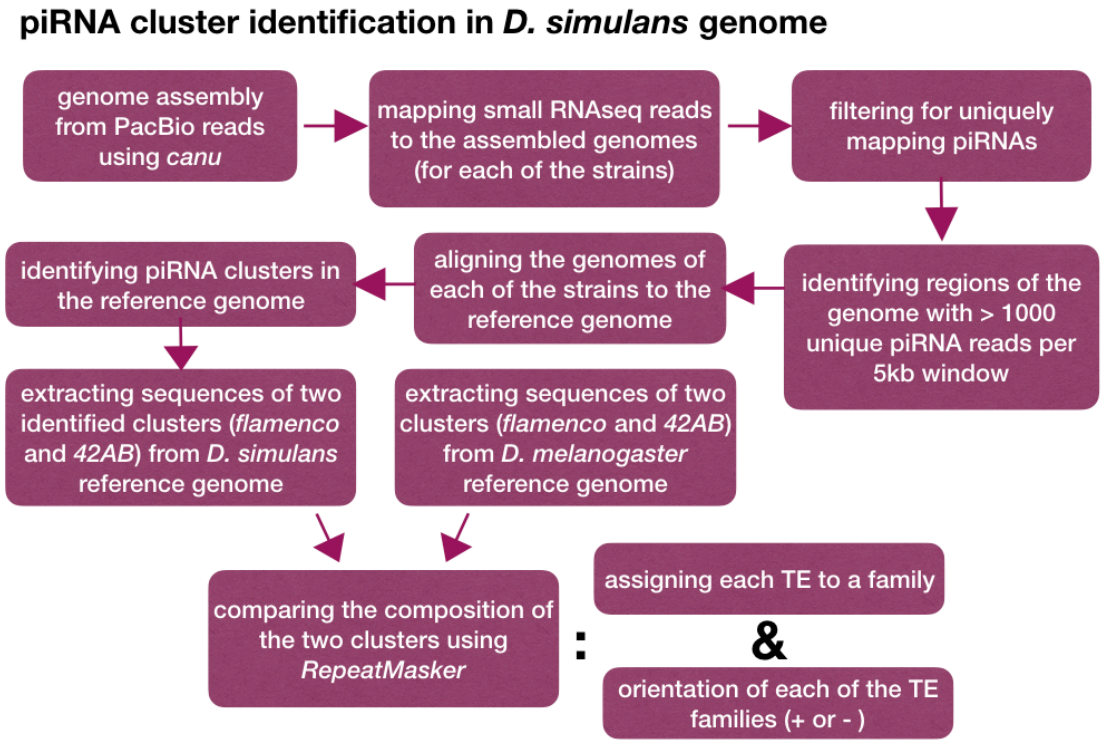


Diagram 4.1 The diagram illustrates the workflow for piRNA cluster identification in the reference genome of *D. simulans* and comparison of the two clusters between *D. simulans* and *D. melanogaster* reference genomes.

4.3 RESULTS

4.3.1 Genome assembly

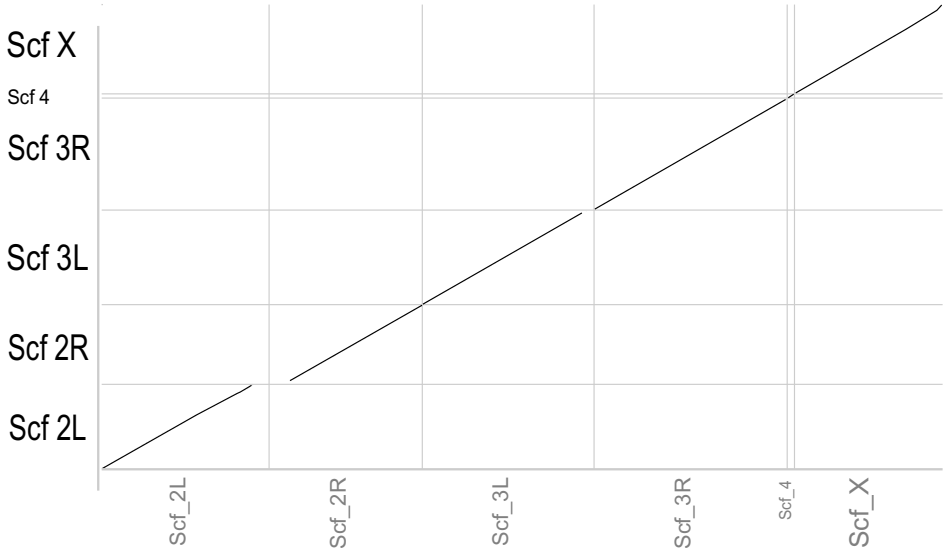
I assembled genomes of two sequenced lines *D. simulans* separately using long reads (Supplementary figure 4.1) obtained from PacBio sequencing. Assembly statistics for each of the strains is shown in table 4.1.

Line	**Lps5**	**SGA26**
Total number of scaffolds	629	486
Total bases	148.355 Mb	147.685 Mb
N/L50	10/3.095 Mb	7/5.953 Mb
Max scaffold length	17.768 Mb	24.597 Mb
Number of scaffolds	279	215
Main genome in scaffolds	93.16%	94.49%

Table 4.1. Assembly statistics for each of the sequenced lines, Lps5 and SGA26.

I compared our assemblies to the existing reference genome from Flybase (dsim_r2.02_FB2017_04). The genomes are colinear across the chromosomes (Fig. 4.1).

A



B

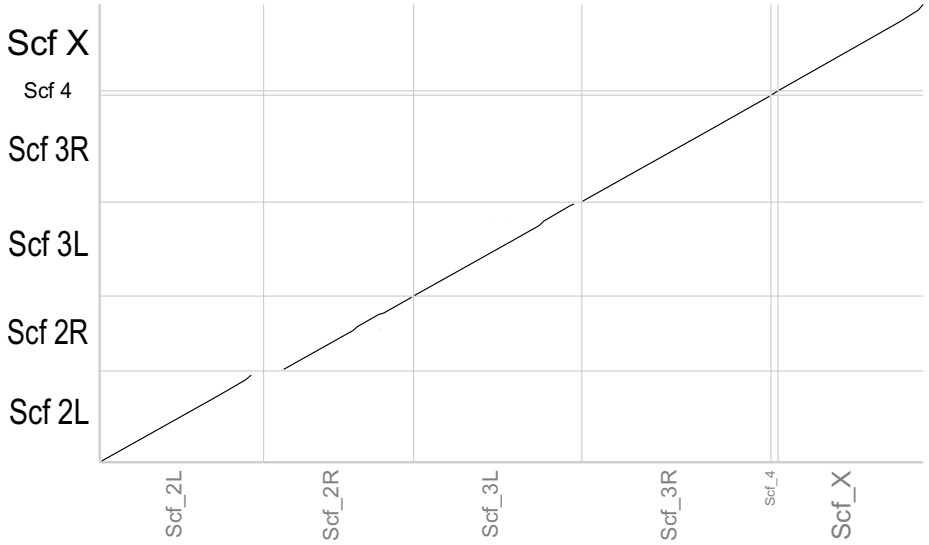
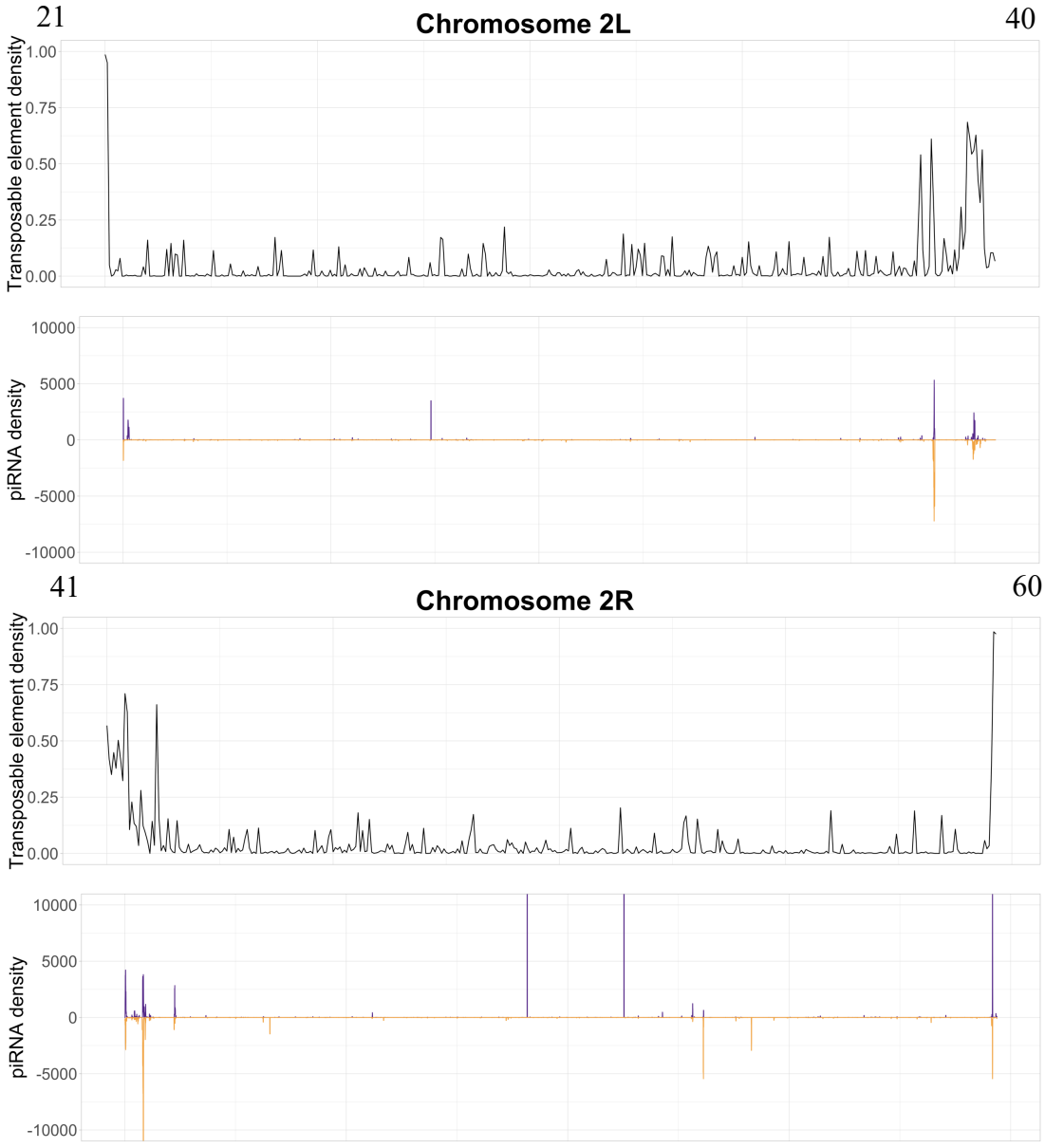


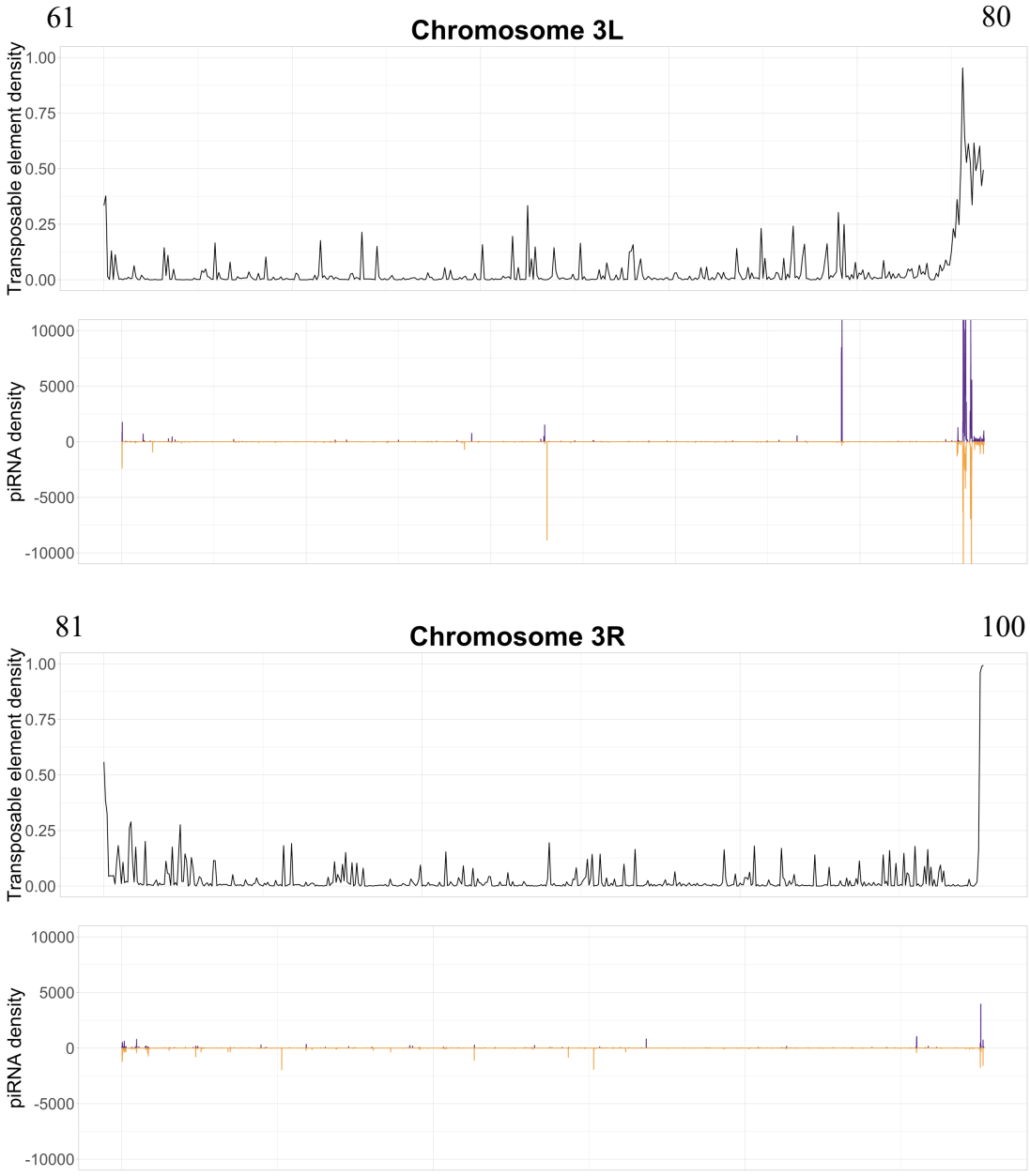
Figure 4.1. Alignment of the *D. simulans* reference from Flybase and (**A**) Lps5 and (**B**) SGA26 assemblies.

4.3.2 piRNA cluster identification

I mapped small RNA sequencing reads obtained from previous experiments (Chapter 2) for each of the lines to the assembled genomes. I restricted my analysis to the uniquely mapping piRNAs, which account for ~18% of the reads (Supplementary table 4.1). I calculated the density of the transposable elements in the assembled genomes per 50kb. By inspection, uniquely mapping piRNAs cluster in the regions of the genome with high TE density (Fig. 4.2 and 4.3).

Figure 4.2. Transposable element density per 50 kb in the assembled genome of Lps5 line and piRNA mapping density (uniquely mapping) in the assembled genome of Lps5 line in 5kb sliding windows (purple: forward, orange: reverse) for each of the chromosomes. Numbers at the top of each plot indicate cytological location.




61
Chromosome 3L
80
Transposable element density
1.00
0.75
0.50
0.25
0.00
piRNA density
10000
5000
0
-5000
-10000
81
Chromosome 3R
100
Transposable element density
1.00
0.75
0.50
0.25
0.00
piRNA density
10000
5000
0
-5000
-10000

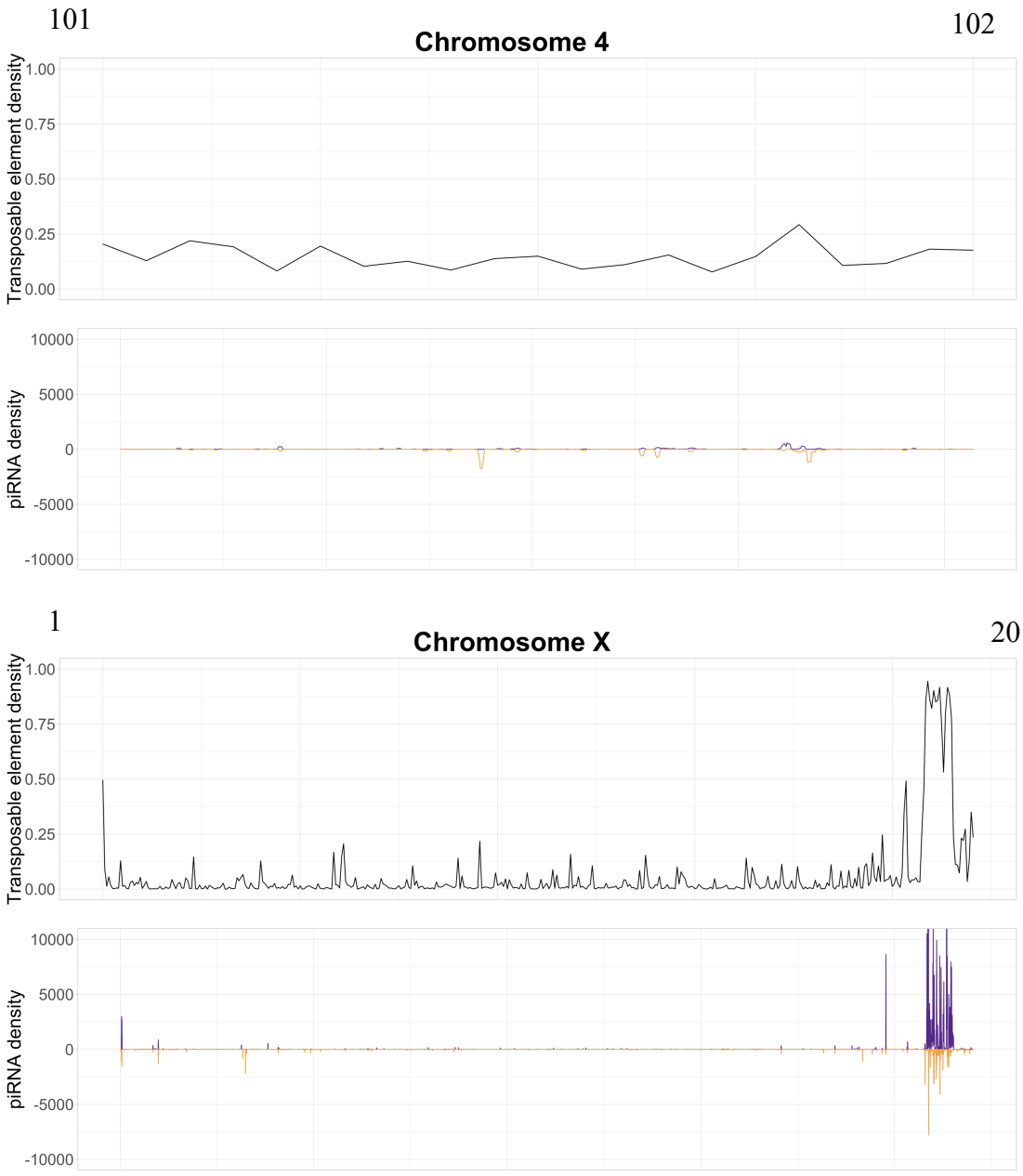
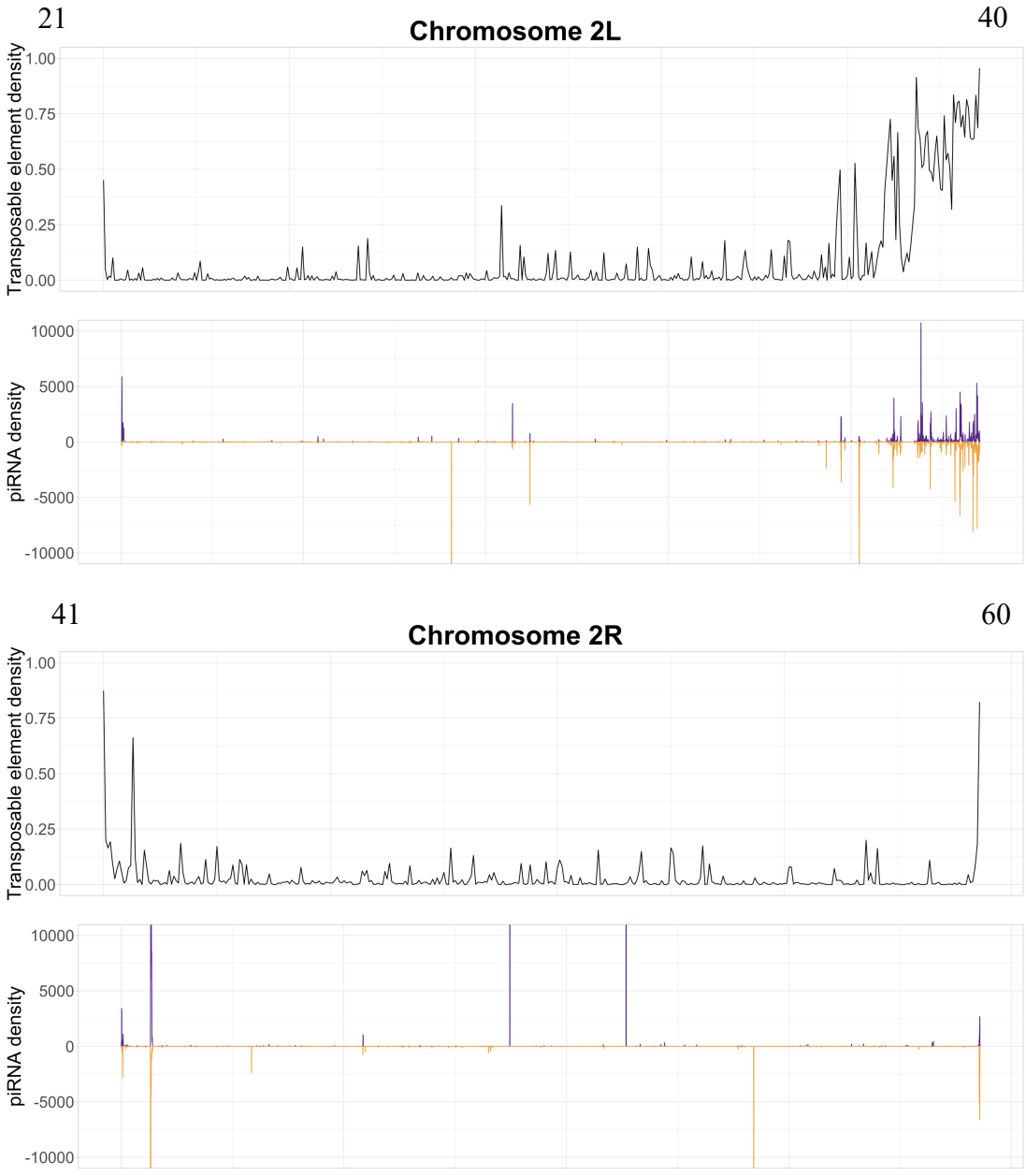
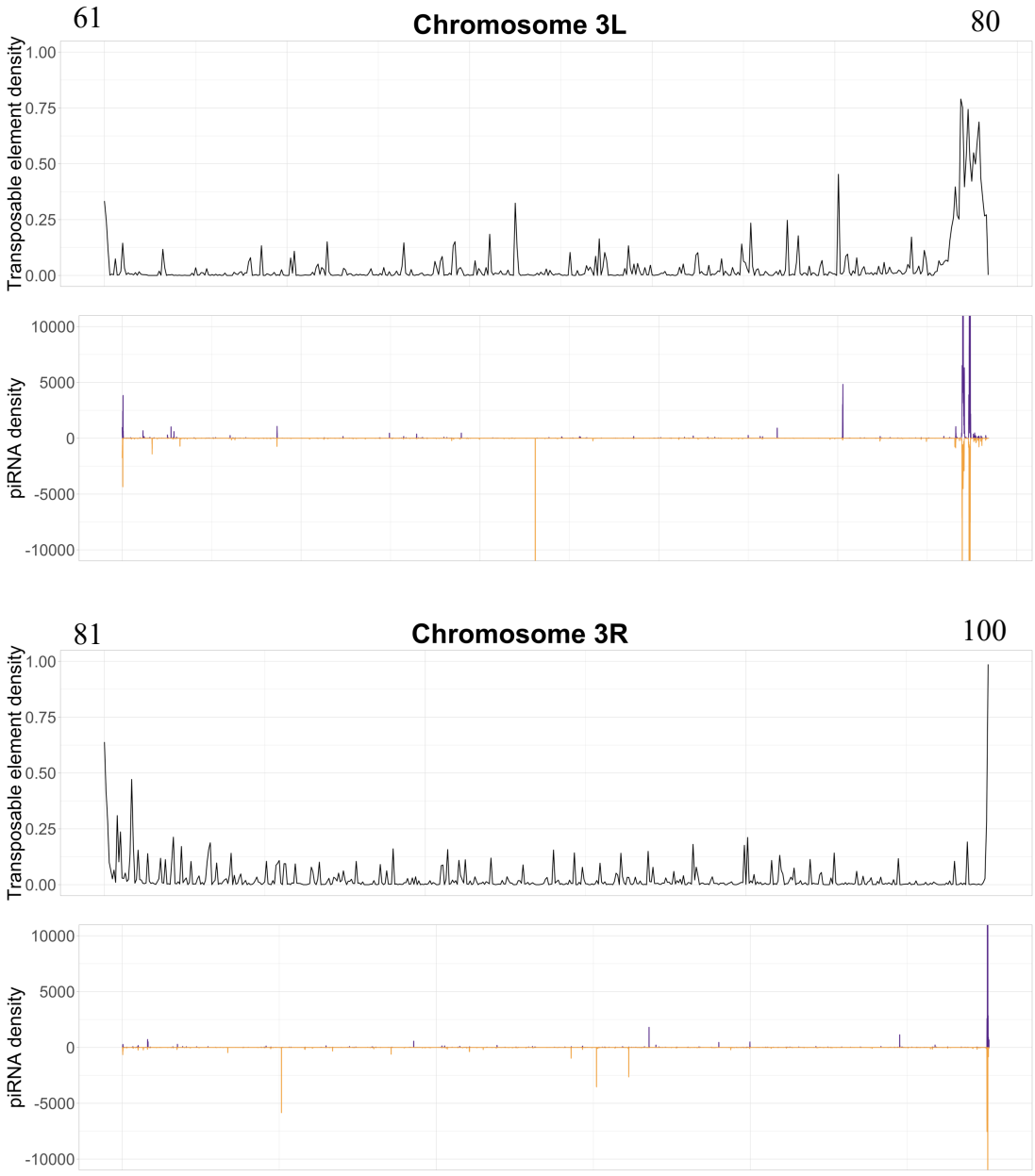
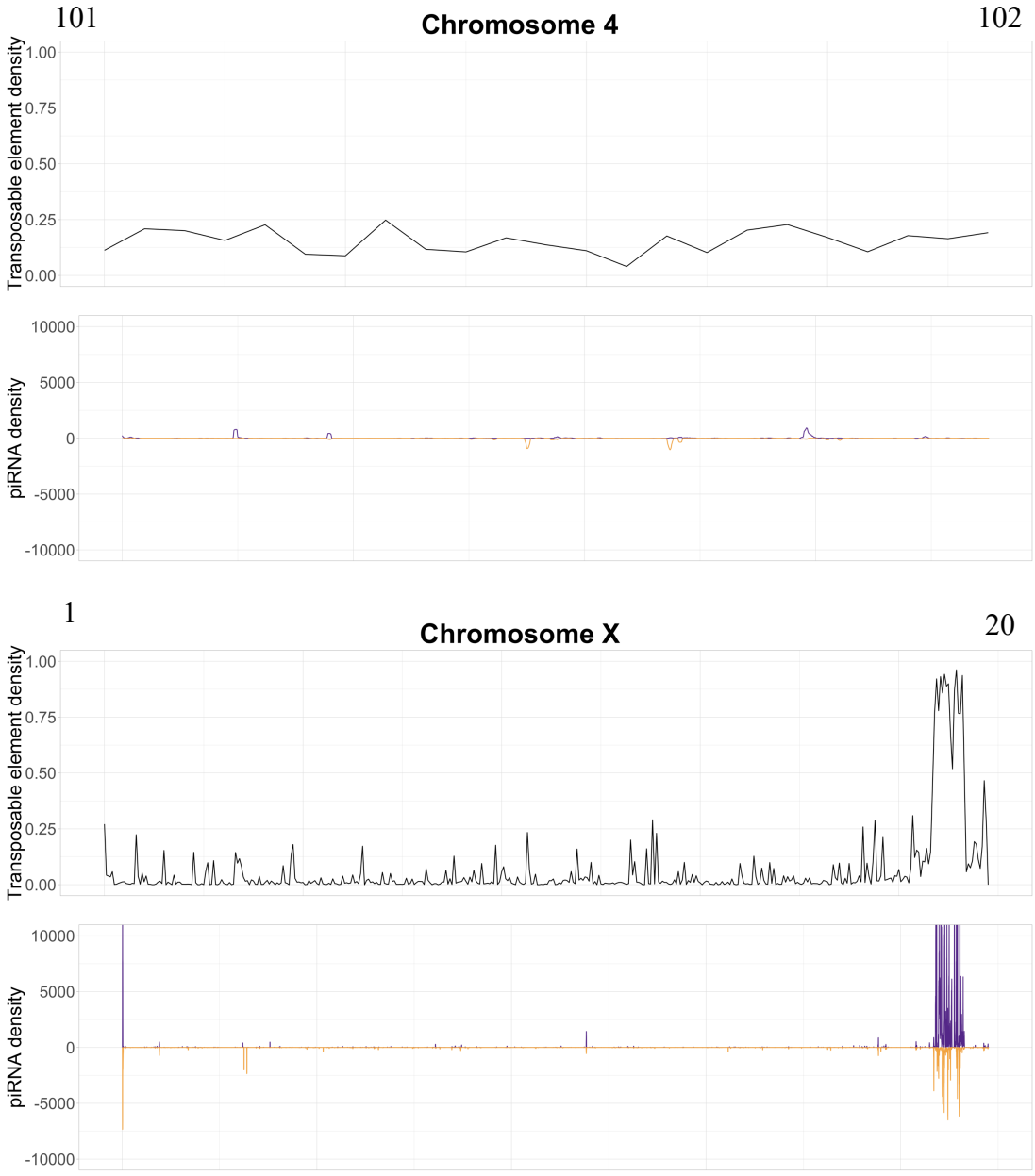

101
102
Chromosome 4
Transposable element density
1.00
0.75
0.50
0.25
0.00
piRNA density
10000
5000
0
-5000
-10000
1
20
Chromosome X
Transposable element density
1.00
0.75
0.50
0.25
0.00
piRNA density
10000
5000
0
-5000
-10000

Figure 4.3. Transposable element density per 50 kb in the assembled genome of SGA26 line and piRNA mapping density (uniquely mapping) in the assembled genome of SGA26 line in 5kb sliding windows (purple: forward, orange: reverse) for each of the chromosomes. Numbers at the top of each plot indicate cytological location.




61
Chromosome 3L
80
Transposable element density
1.00
0.75
0.50
0.25
0.00
piRNA density
10000
5000
0
-5000
-10000
81
Chromosome 3R
100
Transposable element density
1.00
0.75
0.50
0.25
0.00
piRNA density
10000
5000
0
-5000
-10000


101
Chromosome 4
102
Transposable element density
1.00
0.75
0.50
0.25
0.00
piRNA density
10000
5000
0
-5000
-10000
1
Chromosome X
20
Transposable element density
1.00
0.75
0.50
0.25
0.00
piRNA density
10000
5000
0
-5000
-10000

Cluster ID	Chromosome	Start	Stop	Length	TE bases on + strand	TE bases on - strand	Number of TEs on + strand	Number of TEs on – strand	TE % on + strand	TE % on - strand	Present in Lps5	Present in SGA26	piRNA distribution (+/-) in Lps5	piRNA distribution (+/-) in SGA26
1	X	6000	23008	17008	3688	1595	21	9	21.7	9.38	Yes	Yes	67/33	33/67
2	X	859966	866477	6511	136	0	1	0	2.1	0.00	Yes	Yes	40/60	39/61
3	X	19208449	19231954	23505	0	604	0	7	0.0	2.57	Yes	No	93/7	-
4	X	19736655	19783874	47219	10586	6691	24	24	22.4	14.2	Yes	Yes	83/17	71/29
5	X	20177515	20352215	174700	52008	75454	130	134	29.8	43.2	Yes	Yes	93/7	88/12
6	X	20707375	20829444	122069	13805	14857	74	83	11.3	12.2	Yes	Yes	42/58	58/42
7	2L	1	52000	51999	1904	1745	11	13	3.7	3.4	Yes	Yes	97/3	96/4
8	2L	11053898	11061903	8005	110	0	1	0	1.37	0.00	No	Yes	-	13/87

Cluster ID	Chromosome	Start	Stop	Length	TE bases on + strand	TE bases on - strand	Number of TEs on + strand	Number of TEs on – strand	TE % on + strand	TE % on - strand	Present in Lps5	Present in SGA26	piRNA distribution (+/-) in Lps5	piRNA distribution (+/-) in SGA26
9	2L	19558215	19565213	6998	3109	1983	3	10	44.4	28.3	No	Yes	-	49/51
10	2L	20,044,087	20,060,914	16827	10454	1240	32	7	62.1	7.4	Yes	Yes	22/78	4/96
11	2L	20,541,248	20548943	7695	1037	1728	7	7	13.5	22.5	No	Yes	-	18/82
12	2L	20594547	20893880	299333	43012	44796	206	190	14.4	15.0	Yes	Yes	53/47	44/56
13	2L	20946700	21072070	125370	17502	11241	72	56	14.0	9.0	Yes	Yes	32/68	52/48
14	2L	21392273	21574886	182613	35965	53243	145	162	20.0	29.2	No	Yes	-	70/30
15	2R	2219212	2479708	260496	55221	54690	209	241	21.2	21.0	Yes	Yes	29/71	55/45
16	2R	3138469	3257091	118622	22048	54690	89	83	18.6	17.4	Yes	Yes	66/34	40/60

Cluster ID	Chromosome	Start	Stop	Length	TE bases on + strand	TE bases on - strand	Number of TEs on + strand	Number of TEs on – strand	TE % on + strand	TE % on - strand	Present in Lps5	Present in SGA26	piRNA distribution (+/-) in Lps5	piRNA distribution (+/-) in SGA26
17	2R	21462271	21,532,301	70030	1813	1848	10	16	2.59	2.64	Yes	Yes	64/36	26/74
18	3L	1	17500	17499	1068	3242	6	8	6.10	18.5	Yes	Yes	60/40	63/37
19	3L	19346297	19354628	8331	141	5753	1	15	1.69	69.1	Yes	Yes	98/2	97/3
20	3L	22438430	22518216	79786	15966	15417	80	81	20.0	19.3	Yes	Yes	47/53	45/55
21	3L	22687784	22726191	38407	9566	8836	38	41	25.0	23.0	Yes	Yes	72/28	65/35
22	3L	22987876	23006594	18718	2302	12222	12	28	12.3	65.3	Yes	Yes	47/53	47/53
23	3L	23103134	23244423	141289	54768	44166	152	111	38.8	31.3	Yes	Yes	36/64	50/50
24	3L	23971655	24153931	182276	49723	55441	140	159	27.3	30.4	Yes	Yes	51/49	38/62

Cluster ID	Chromosome	Start	Stop	Length	TE bases on + strand	TE bases on - strand	Number of TEs on + strand	Number of TEs on – strand	TE % on + strand	TE % on - strand	Present in Lps5	Present in SGA26	piRNA distribution (+/-) in Lps5	piRNA distribution (+/-) in SGA26
25	3R	1	28,653	28652	5165	2277	28	14	18.0	7.95	Yes	Yes	56/44	36/64
26	3R	327692	373374	45682	4990	5560	31	28	11.0	12.2	Yes	Yes	37/63	85/15
27	3R	692374	706848	14474	1406	440	14	3	9.71	3.04	Yes	Yes	80/20	26/74
28	3R	27136820	27160941	24121	1210	296	9	2	5.02	1.23	Yes	Yes	69/31	26/74

Table 4.2. Major piRNA clusters identified in *D. simulans* genome (Flybase ftp://ftp.flybase.net, dsim_r2.02_FB2017_04). piRNA strand distribution shows the percentage of piRNAs mapping to + and – strands in each of the assembled genomes. Only piRNAs mapping uniquely to the genome were taken into account to calculate piRNA strand distribution. Cluster IDs displayed in red indicate clusters that are present in only one of the lines examined.

Using my own criterion of 1000 uniquely mapping piRNAs being found within a 5kb window, I identified 28 major piRNA clusters in *D. simulans* (Table 4.2).

Most clusters (23) are present in both of the sequenced strains of *D. simulans*, whereas five of them appear to exist in one of the strains but not the other. *De novo* formation of the clusters in one of the strains could explain the observed differences in the cluster presence. In *D. melanogaster,* transgenes inserted in euchromatic regions of the genome showed to initiate *de novo* piRNA cluster formation and small RNA synthesis (Olovnikov *et al.*, 2013). One of the clusters (ID:14, chromosome 2L) is not present in Lps5 due to the shorter length of the Lps5 chromosome assembly, which does not cover the cluster region.

As in *D. melanogaster*, piRNA clusters identified in *D. simulans* vary in size and some of them are as long as 260 kb in the reference genome. By inspection, most of the clusters occur in heterochromatic regions of the genome – in centromeres and telomeres.

In both assemblies, I identified two clusters on the X chromosome, and both of these clusters occur in the genomes assembled here. However, these align to one location in the reference genome (cluster 5, table 2). This could be a result of either duplication of the region in both of the strains or assembly artefacts.

4.3.3 piRNA cluster comparison with *D. melanogaster*

Flamenco

One of the most studied clusters in *D. melanogaster* is *flamenco*, known to control the retrotransposons – *gypsy*, *ZAM* and *Idefix* (Pelisson *et al.*, 1994; Desset *et*

al., 1999; Desset *et al.*, 2003; Goriaux *et al.*, 2014). This cluster, located downstream from the *DIP-1* gene, is a uni-directional cluster that controls transposons in the somatic support cells of the *Drosophila* germline (Brennecke *et al.*, 2007). Transposable elements in *flamenco* are inserted in antisense orientation and piRNAs produced from this cluster are mostly sense (Brennecke *et al.*, 2007). After identifying piRNA clusters in *D. simulans* reference genome, I investigated whether *flamenco* is also present in *D. simulans* by looking at the *DIP-1* gene location. I found that cluster 5 on the X chromosome is proximal to *DIP-1* gene. I compared TE classes present in *flamenco* in both *D. melanogaster* and *D. simulans* and their orientation (Fig. 4.4).

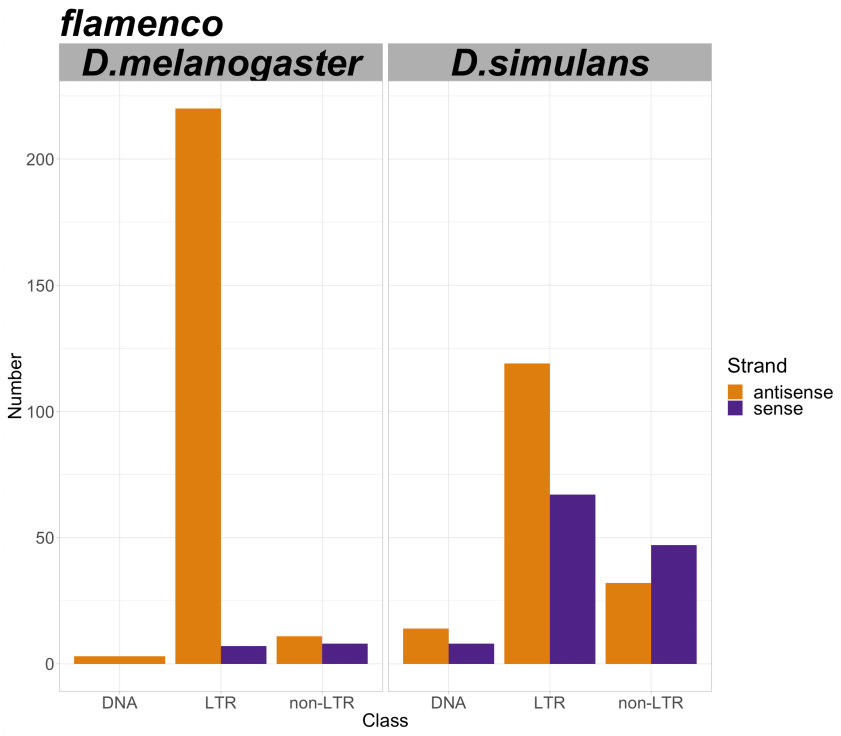


Figure 4.4. Classes of TEs present at *flamenco* locus in *D. melanogaster* and *D. simulans*. Colour of the bars indicates the TE orientation in the cluster (purple: sense, orange: antisense).

Unlike in *D. melanogaster*, transposable elements in the *D. simulans flamenco* locus do not seem to be inserted in one orientation. In both of the assembled genomes, two clusters that are located next to each other align to the *flamenco* cluster in the reference genome. I looked at the TE content for each of these clusters (called 5a and 5b) in each of the strains (Fig. 4.5). The distribution of piRNAs mapping uniquely to each of the clusters in both strains is sense biased in both lines (forward/reverse percentage of the piRNAs in the 5a cluster: 81/19 and 85/15 in Lps5 and SGA26, respectively; 92/8 and 87/13 for cluster 5b in Lps5 and SGA26,

respectively). This biased piRNA distribution suggests that piRNAs are generated from one strand of the cluster and these clusters are likely to be uni-directional in both of the strains. However, transposable elements in these clusters are inserted in both orientations, which is unexpected for a uni-directional cluster. The reason is that only the anti-sense piRNAs generated from these clusters will be protective.

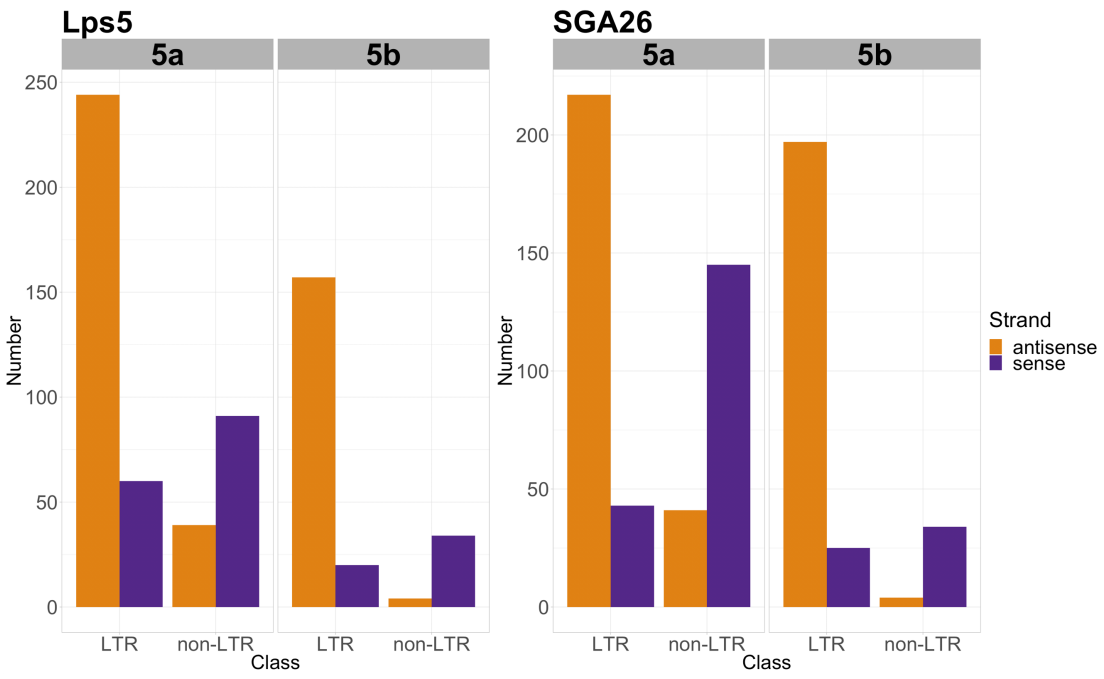


Figure 4.5. Numbers of TE classes in clusters 5a and 5b in the assembled genomes. Both of the clusters correspond to cluster 5 in the reference genome. Colours represent the orientation of the TEs (purple: sense, orange: antisense).

42AB

One of the biggest (240 kb) bi-directional clusters in *D. melanogaster* is *42AB* located on chromosome 2R. It is known to control most of the TEs in the genome (Brennecke *et al.*, 2008). I found a ~260 kb cluster (cluster ID15, table 2) on chromosome 2R in *D. simulans*. Transposable element orientation within the cluster and piRNA mapping distribution indicates this cluster is bi-directional. As in *D. melanogaster*, cluster 42AB in *D. simulans* contains TEs from all three classes (Fig. 4.6).

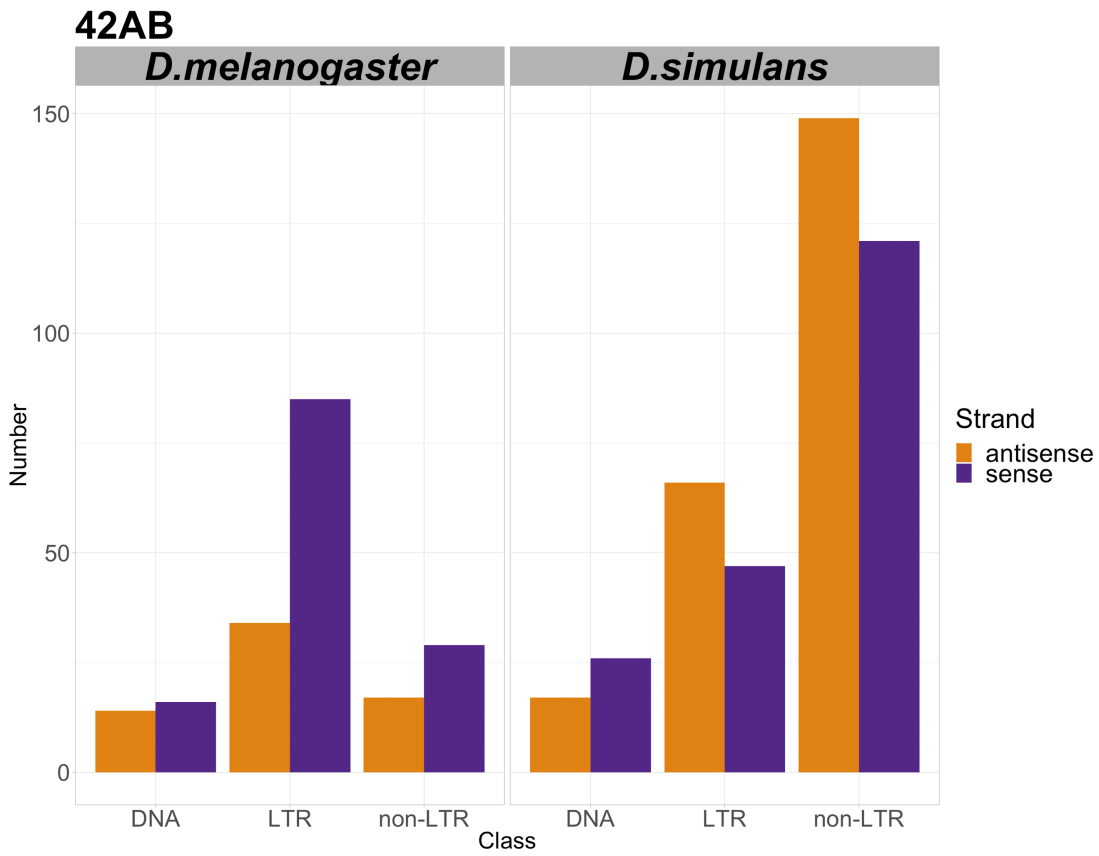


Figure 4.6. Numbers of transposable elements classes within *42AB* piRNA cluster in *D. melanogaster* and *D. simulans* reference genomes. Colour of the bars indicates orientation (purple: sense, orange: antisense).

4.4 DISCUSSION

I present a genome assembly of two *D. simulans* lines from long-read sequencing. I then use small RNA sequencing from the same lines to identify piRNA clusters in the assembled genomes. By aligning the assembled genomes to the reference *D. simulans* genome, I then determine the location of 28 piRNA clusters in the reference genome (Flybase ftp://ftp.flybase.net, dsim_r2.02_FB2017_04).

By inspection, most clusters appear in pericentromeric and telomeric regions of the genome. These heterochromatic clusters are found in the same position in both of the genomes. Some clusters are found in one of the strains, but not the other. One of these clusters is simply not present in the assembly of one of the strains due to a shorter chromosome assembly (Cluster ID 14). Other clusters, present in only one of the strains, mostly do not exceed 10kb in length and by inspection seem to occur in euchromatic regions of the genome. One potential explanation for the such differences in the cluster presence between the lines is *de novo* cluster formation due to a transposable element insertions in the region in one of the lines. It has been shown that in *D. melanogaster*, insertions of transgenes in euchromatic regions of the genome trigger *de novo* piRNA cluster formation and small RNA expression (Olovnikov *et al*., 2013).

I compare two of the major piRNA clusters described in *D. melanogaster* – *flamenco* and *42AB* – to the orthologous loci in *D. simulans* and find piRNA-producing loci in the same regions of the genome. For 42AB, that is a major bi-directional piRNA cluster controlling different classes of TEs in *D. melanogaster*, I find that in *D. simulans* reference genome this cluster also contains all three classes of TEs inserted

in both orientations. From piRNA strand distribution originating from the cluster in both of the strains, it is likely that cluster *42AB* in *D. simulans* is producing piRNAs from both genomic strands and is therefore bidirectional.

The *flamenco* locus in *D. melanogaster* is a uni-directional cluster containing mostly LTR transposons in an antisense orientation In *D. simulans*, this cluster seems to have only a slightly biased antisense transposable element orientation, but the piRNA strand distribution mapping to *flamenco* in *D. simulans* assembled genomes (Lps5 and SGA26) suggests that this cluster is also uni-directional in *D. simulans*. Single molecule RNA fluorescent hybridisation would be helpful to confirm the strandedness of the cluster. In this case, RNA FISH probes for sense and antisense transcripts of the cluster can be used. Similar to *D. melanogaster*, *flamenco* in *D. simulans* consists mostly of retrotransposons. Unlike in *D. melanogaster*, where mostly LTRs insertions are found, *D. simulans flamenco* contains both LTR and non-LTR classes.

piRNA clusters control transposable elements in the germline; identifying these piRNA producing loci in *D. simulans* will give an opportunity to study the evolution of these important component of transposable elements defence. By looking at the TE insertions in the clusters in different lines, it would be possible to see how many insertions of each of the elements the clusters have and whether the insertions happen to be in the same place of the clusters or different. This comparison may give an idea about the evolution of the clusters – insertion in one place in all of the clusters would suggest a single insertion event that has spread in the populations; different insertions in the lines would be an indication of several

insertions in the ancestral population. Another interesting question would be investigating the spatial and temporal transcription of the clusters – whether all of them are active at the same time in all of the cells of the same type.

Chapter 5 Discussion

Selfish genetic elements have been widely studied since their discovery. Transposable elements, being the most taxonomically widespread example of a selfish genetic element, have been a focus of research for several reasons. Transposons can be used as a tool for genetic manipulations of *Drosophila* and other species. In addition, a large proportion of the human genome is composed of transposable elements, and their occasional mobilisation leads to some genetic disorders.

Major defence against transposable elements in the germline of animals involve piRNAs, small non-coding RNAs that are encoded by a small number of genomic loci. This knowledge comes from studies on transposable elements and host defence systems against them have been focused on transposable elements that have invaded and spread within a species (Brennecke *et al*., 2007). However, little is known about the mechanisms involved in TE defence during the early phases of TE invasion. One of the main aims of this thesis was to unravel the host mechanisms involved in the defence against one of the most studied transposable elements in *Drosophila*, the *P*-element, during early phases of invasion of this element in *D. simulans*.

Chapter 2 investigates whether small non-coding RNAs, piRNAs, are the most important factor protecting the host germline against *P*-element activity during early stages of invasion. I used *D. simulans* lines collected during initial stages of the *P*-element invasion, which exhibit variation in tolerance to the negative consequences

of uncontrolled *P*-element transposition. piRNAs are the most important factor in protecting the germline against established transposable elements (Brennecke *et al*., 2007), but I show that they are unlikely to be acting during the early stages of this transposable element invasion.

Chapter 3 is aimed at understanding the genetic architecture of tolerance to HD: how many loci underlie variation in tolerance, and whether these loci are dominant or recessive. Using a set of crosses, I looked at the inheritance patterns of tolerance. My results suggest that tolerance is a quantitative trait with likely more than one locus, neither fully recessive not fully dominant.

Chapter 4 characterises genomic piRNA clusters in *D. simulans*. piRNA clusters are heterochromatic locations of the genome that consist of nested transposable elements and serve as piRNA-producing loci. I used small RNA sequencing and long-read genome sequencing data to identify piRNA clusters in *D. simulans*. I identified 28 clusters in the *D. simulans* genome and compared the composition of two major clusters between *D. melanogaster* and *D. simulans*.

The hypothesis that mechanisms of defence against established within the species transposable elements (e.g. piRNAs) are different from the ones involved in the early stages of invasion of a TE is perhaps the most interesting finding of this thesis. The process of establishing defence against transposable elements can be compared to the immune response. In the initial stages of invasion of a transposable element, mechanisms not specific to this transposon act to eliminate the negative consequences of transposition to ensure cell survival ('innate' immune response). In the later stages of a transposon invasion, piRNAs target transposable elements

transcripts based on sequence complementarity, and are produced only against transposable elements that are already present in the genome. Therefore, the piRNA pathway can be compared to 'adaptive' immune response.

One of the possible explanations for the variation in P tolerance in the absence of piRNAs is epigenetic suppression, TE silencing might be more efficient in more tolerant lines of *D. simulans* via chromatin modifications (Lee and Karpen, 2017). Uncontrolled transposition is a source of double-stranded DNA breaks (DSB) that are toxic to the cell and, if not repaired, can lead to apoptosis, programmed cell death. It is possible that tolerance to the *P*-element-induced HD is not specific to the *P*-element, but a more general mechanism that allows cells to deal with stress. More efficient DNA repair in the more tolerant lines can cause differences in tolerance to the *P*-element-induced damage. There is variation in tolerance to UVB in *Drosophila* that is positively correlated with expression of DNA damage response genes (Svetec *et al*., 2016). Alternatively, as hybrid dysgenesis is a consequence of apoptosis, genes that regulate apoptosis may also provide tolerance to *P*-element-induced damage, independent of piRNAs. Candidates for this include the genes *p53* and *bruno*, both of which are involved in non-piRNA dysgenesis tolerance in *D. melanogaster* (Kelleher *et al.,* 2018; Tasnim and Kelleher, 2018). If these tolerance genes play a general role in coevolution between hosts and their TEs, we might expect to see signatures of positive selection of these genes. In fact, several genes involved in the maintenance of female germline stem cells show signs of positive selection (Kelleher *et al.,* 2018).

Tolerance to *P*-element-induced damage may be an example of the robustness of the system (Kitano, 2014). The cells are able to maintain their

functionality when experiencing mutation or stimuli (Kitano, 2014). In the cells, there are several ways to achieve the same function, so that the failure of one does not lead to the failure of the whole systems and therefore ensures cell survival, so called 'heterogeneity' (Kitano, 2014).

I used a set of crosses to estimate the number of loci involved in P tolerance. P tolerance seems to be a polygenic trait with the alleles involved in it being not fully recessive or fully dominant. Standard QTL mapping techniques, e.g. an advanced intercross mapping method, should be used to get a better estimation of the genetic basis of tolerance.

It would be interesting to introduce another new transposable element into the same lines and see whether the lines are also tolerant to this new TE, to which the species are still naïve. Whether tolerance is a result of a more efficient DNA repair system in the lines, might be investigated by testing these flies' ability to repair DSBs introduced by UV exposure.

Tolerance to a TE is a trait that, on the one hand, protects the host and allows for the organisms to be robust to different stresses. On the other hand, it allows a transposable element to invade a population and spread within it, as it masks the negative consequences of transposition that, if being extremely deleterious, would lead to sterility or death of a host, therefore preventing the spread of a TE. It could be that tolerance is a trait that determines whether a transposon can invade and spread in the species. On an organismal level, being tolerant to a TE allows the organism to survive, but on a population level, lack of tolerance is one of the ways to stop TE invasion and spread. Tolerance could be a factor that effects the rate of TE

spread. If so, there may have been previous attempts of the *P*-element to invade *D. simulans* that were not successful due to host factors.

Many questions still remain in the study system examined in this thesis. Does the host or the TE need to be pre-adapted for an invasion? If so, what do these pre-adaptations include in both the host and the TE? We only see successful attempts of a TE invasion – how many attempts of an invasion are unsuccessful due to high cost imposed on the host? Is tolerance caused by the same factors in all of the tolerant individuals or can it be achieved by different means? Examining tolerance in different populations and species may give an insight into this question. Further, why do some flies lack tolerance? Tolerance might be a costly trait, or the least tolerant flies might benefit from their lack of tolerance in an unknown way.

Another important question is how TE invasions of a new species happen? TEs can be horizontally transferred between species, and the *P*-element is the most recent example of a transfer between two sister species – *D. melanogaster* and *D. simulans*. However, it is not clear by which means a DNA cut-and-paste transposon was introduced to *D. simulans*, most likely coming from *D. melanogaster*. Unlike retrotransposons, that can be encapsulated and form virus-like particles, DNA transposons do not spread in virus-like manner. A retroelement *gypsy* can invade the germline of *D. melanogaster* when larvae are grown on a medium containing enveloped gypsy particles. Could it be possible that *P*-element was introduced into *D. simulans* by a similar mechanism, being encapsulated along with some other retroelement or a virus?

In the germline, piRNA clusters control transposable elements. Identification of piRNA clusters in *D. simulans* may give an insight into the evolution of adaptive TE defence in these species. By comparing piRNA cluster composition it might be possible to see whether the insertions of the TEs into the clusters are shared between different strains and how conserved the clusters are between the species. Investigating the temporal and spatial transcription of the clusters, whether all of the clusters are active at the same time, would be another interesting question.

Supplementary information

Chapter 2

Supplementary table 2.1

Primer sequences for the amplification of *P*-element exons

Primer	**Sequence**
P-element exon 0 Forward	GGTTGTGTGCGGACGAATTTT
P-element exon 0 Reverse	CTGGTTCAGGCTCTATCACTTT
P-element exon 1 Forward	TCTACGCAAAATCTTCACGGAC
P-element exon 1 Reverse	CTGATATACCGAGCTCTGTCCA
P-element exon 2 Forward	TCCTGCAGATGACCATTTAAAGA
P-element exon 2 Reverse	TTAAACTGCAGTGGAGTGGGAT
P-element exon 3 Forward	GGACAACTCTGAAAGCTGGC
P-element exon 3 Reverse	CGTTTCGCGCTGCTAATATTAA

Supplementary table 2.2

Primer sequences for *hobo* (efficiency 1.94) and *rp49* (efficiency 1.96) qPCR

Primer	**Sequence**
Hobo forward	AGGGCAATACCCGTTTAAATTGT
Hobo reverse	AGGCGCTTTTCAAAGTGGTTT
Rp49 forward	CGGATCGATATGCTAAGCTGT
Rp49 reverse	GCCCTTGTTCGATCCGTA

Supplementary table 2.3

Primer sequences for *P*-element spliced (efficiency 1.98) and non-spliced (efficiency 1.94) qPCR and *rp49* qPCR (efficiency 1.96)

Primer	**Sequence**
P-element spliced F	GTATAGGTTAAGAAAATATATAATAGCCA
P-element spliced R	TCATCGACAGGCTCATCATC
P-element overall F	TGAGTGCTCGCAACCTTATG
P-element overall R	GCCATCAAGCGAAGCATTAT
Rp49 forward	CGGATCGATATGCTAAGCTGT
Rp49 reverse	GCCCTTGTTCGATCCGTA

Supplementary table 2.4

Presence of *P*-element exons in the genome of the studied lines as measured by PCR; primers used are in Supplementary Table 1.

Line	**Exon 0**	**Exon 1**	**Exon 2**	**Exon 3**
SGA01	x		x	x
SGA02	x		x	x
SGA09	x	x	x	x
SGA11	x			x
SGA12	x		x	
SGA13	x			x
SGA14	x	x		x
SGA15	x			x
SGA16	x	x	x	x
SGA17	x	x	x	x
SGA18	x			x
SGA20	x			x
SGA22				
SGA24	x	x	x	x
SGA26	x		x	x
SGA27	x		x	x
SGA32	x	x	x	x
SGA33	x	x	x	x
SGA34	x			
SGA35	x	x	x	x
Lps1	x			x
Lps2	x			x
Lps3	x			x
Lps5	x			x
Lps6	x			x

Lps12	x			x
Lps13	x			x
Hin17				

Supplementary table 2.5

Number of raw reads obtained per library

Line	Number of reads replicate 1	Number of reads replicate 2
Hin17	9 533 199	11 201 041
Lps3	10 114 116	11 241 463
Lps5	10 828 654	10 501 044
Lps6	11 267 174	11 140 035
SGA02	11 415 313	10 670 261
SGA12	20 070 344	5 370261
SGA14	10 605 511	18 712 295
SGA18	10 038 313	11 356 245
SGA20	11 980 299	11 346 904
SGA26	10 484 350	10 158 381
SGA27	11 735 971	12 148 053
SGA34	11 005 258	10 596 808

Supplementary table 2.6

Number of raw reads obtained per library for the second round of small RNA sequencing

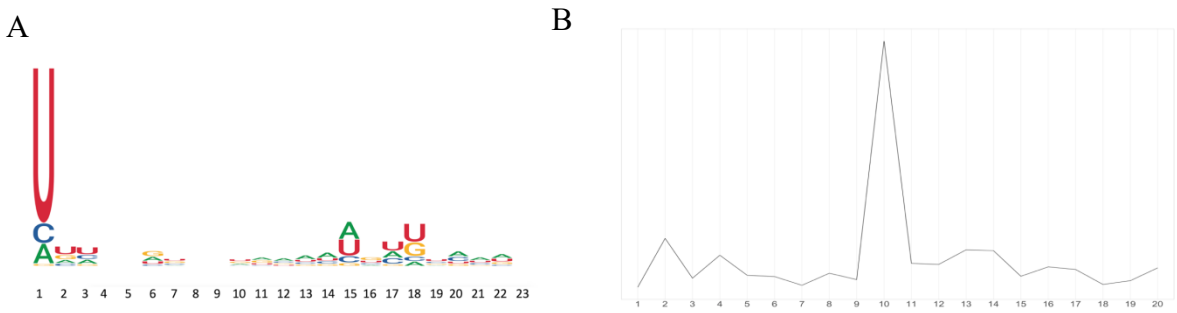
Line	Number of reads replicate 1	Number of reads replicate 2
Cro18	13 010 655	12 136 578
Lps5	12 128 019	13 641 904
SGA14	16 175 037	17 925 604
SGA26	17 142 273	17 599 985
SGA27	13 805 877	11 324 225
Lps5xCro18 (F1)	14 202 189	15 365 892
SGA14xCro18 (F1)	15 123 361	16 936 916
SGA26xCro18 (F1)	18 505 783	10 247 106
SGA27xCro18 (F1)	16 779 440	13 380 621

Supplementary table 2.7

p-values of binomial exact test for P tolerance of F1 offsring having tolerance of least, most tolerant parent or an intermediate tolerance of the two parents

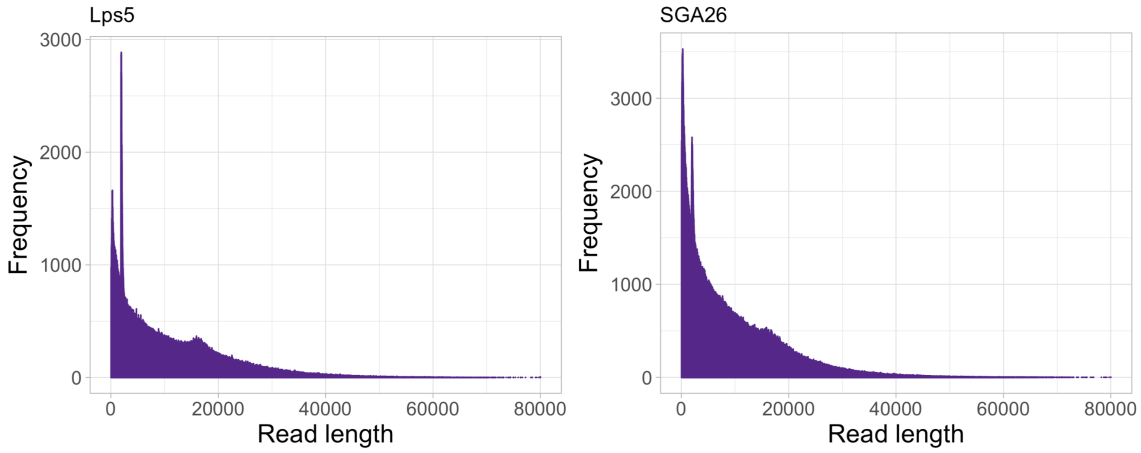
Cross	*p*-value least tolerant parent	*p*-value most tolerant parent	*p*-value intermediate tolerance
SGA14xLps5	1.69E-11	2.20E-16	2.20E-16
Lps5xSGA14	7.77E-08	2.20E-16	2.20E-16
SGA26xLps5	0.6138	2.20E-16	2.20E-16
Lps5xSGA26	1	2.20E-16	5.28E-13
SGA14xSGA27	1.72E-06	2.20E-16	3.57E-15
SGA27xSGA14	4.59E-05	2.20E-16	3.49E-13
SGA26xSGA27	0.3339	2.20E-16	2.20E-16
SGA27xSGA26	0.07353	2.20E-16	2.25E-11

Supplementary figure 2.1. (A) The relative nucleotide bias for each position in piRNA reads starting at 5' end. (B) Graph showing the overlap between forward and reverse piRNA reads. The spike at position 10 indicates that forward and reverse reads overlap at 10 nt. This shows the characteristic ping-pong signal of TEs in the sequenced small RNAs.



Chapter 4

Supplementary figure 4.1. Length distribution of the obtained raw PacBio reads



Supplementary table 4.1. Mapping statistics of the small RNA seq reads to the assembled genomes. The reads were mapping to the assembled genome with 0 mismatches.

Line	Total reads (small RNA seq)	Reads mapped to the assembled genome	Uniquely mapping piRNAs (number/percentage)
Lps5	25 769 923	23 792 923	3 780 334 / 16%
SGA26	34 742 258	32 285 673	5 887 558 / 18%

Supplementary table 4.2. Primers used to amplify 10 genes across *D. simulans* genome. The annealing temperature for all of the PCR reactions was 56 °C.

Primer	**Sequence**
GD10140_1_F	AGCAGTGGAGAGAGCAAGTT
GD10140_1_R	CCAAATCGTCAGCTTCCTCG
GD22574_1_F	GGACTGTGGCACCTCTTACT
GD22574_1_R	GGTAACCAAAGCGCAACTGA
GD24072_1_F	TATATGCGGCATGGTTCCCT
GD24072_1_R	GGTGGCCGATCTTGTTGTTT
ppk6_1_F	CATCTTTGGCATGGACGGAG
ppk6_1_R	TTAGTCCCAGGCAGAGGTTG
GD20943_1_F	AAACCAAATTCCCGCAGTCC
GD20943_1_R	TTAGTCCTTCCTGCAAGCGA
GD20057_1_F	CGGATTTTGGACCTATCGCG
GD20057_1_R	AGGTGTCCTTGGCTAGCATT
GD13128_1_F	GTGTGCCTCATCGATGTCTG
GD13128_1_R	GGTGGGAGTCAAGATCTGCT
GD12511_1_F	TGGGTTTGACTGATTGCACG
GD12511_1_R	GGCTTTGCTGAACCATTGGA
GD16199_1_F	TTCAAGAGCAAACCACAGGC
GD16199_1_R	ACGTCCTGAAAGTTAGCCGA
GD15598_1_F	CAAGCGTAACGTGATCCTGG
GD15598_1_R	ACTTGAATCTCCAGCCACGA

References

Adams, M. D., Tarng, R. S., and Rio, D. C. (1997). The alternative splicing factor PSI regulates P-element third intron splicing in vivo. *Genes Dev, 11*(1), 129-138

Ajioka, J. W., and Eanes, W. F. (1989). The accumulation of *P*-elements on the tip of the X chromosome in populations of *Drosophila melanogaster*. *Genet Res, 53*(1), 1-6

Andrews, J. D., and Gloor, G. B. (1995). A role for the Kp leucine zipper in regulating P element transposition in *Drosophila melanogaster*. *Genetics*, 141, 587-594

Anxolabehere, D., Kidwell, M. G., and Periquet, G. (1988). Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila melanogaster* by mobile P elements. *Mol Biol Evol, 5*(3), 252-269

Anxolabehere, D., Nouaud, D., Periquet, G., and Tchen, P. (1985). P-element distribution in Eurasian populations of *Drosophila melanogaster*: A genetic and molecular analysis. *Proc Natl Acad Sci U S A, 82*(16), 5418-5422

Aravin, A. A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K. F., Bestor, T., Hannon, G. J. (2008). A piRNA pathway primed by individual transposons is linked to *de novo* DNA methylation in mice. *Mol Cell, 31*(6), 785-799

Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G. J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science, 316*(5825), 744-747

Arca, B., Zabalou, S., Loukeris, T. G., and Savakis, C. (1997). Mobilization of a Minos transposon in *Drosophila melanogaster* chromosomes and chromatid repair by heteroduplex formation. *Genetics, 145*(2), 267-279

Arkhipova, I., and Meselson, M. (2000). Transposable elements in sexual and ancient asexual taxa. *Proc Natl Acad Sci U S A, 97*(26), 14473-14477

Ayarpadikannan, S., and Kim, H. S. (2014). The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases. *Genomics Inform, 12*(3), 98-104

Bachmann, A., and Knust, E. (2008). The use of P-element transposons to generate transgenic flies. *Methods Mol Biol, 420*, 61-77

Beall, E. L., and Rio, D. C. (1996). *Drosophila* IRBP/Ku p70 corresponds to the mutagen-sensitive mus309 gene and is involved in P-element excision in vivo. *Genes Dev, 10*(8), 921-933

Béguin, P., Charpin, N., Koonin, E. V., Forterre, P., and Krupovic, M. (2016). *Casposon* integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. *Nucleic Acids Res, 44*(21), 10367-10376

Bejerano, G., Lowe, C. B., Ahituv, N., King, B., Siepel, A., Salama, S. R., Rubin, E. M., Kent, W. J., Haussler, D. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature, 441*(7089), 87-90

Bingham, P. M., Kidwell, M. G., and Rubin, G. M. (1982). The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. *Cell, 29*(3), 995-1004

Black, D. M., Jackson, M. S., Kidwell, M. G., and Dover, G. A. (1987). KP elements repress P-induced hybrid dysgenesis in *Drosophila melanogaster*. *EMBO J, 6*(13), 4125-4135

Blackman, R. K., Grimaila, R., Koehler, M. M., and Gelbart, W. M. (1987). Mobilization of *hobo* elements residing within the decapentaplegic gene complex: suggestion of a new hybrid dysgenesis system in *Drosophila melanogaster*. *Cell, 49*(4), 497-505

Blumenstiel J. (2010). Evolutionary dynamics of transposable elements in a small RNA world. *Trends Genet* 27(1):23-31

Boeke, J., and Stoye, J. (1997). Retrotransposons, endogenous retroviruses, and the evolution of retroelements. *Cold Spring Harbor*

Bourque, G. (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr Opin Genet Dev, 19*(6), 607-612

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell, 128*(6), 1089-1103

Brennecke, J., Malone, C. D., Aravin, A. A., Sachidanandam, R., Stark, A., and Hannon, G. J. (2008). An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science, 322*(5906), 1387-1392

Brookfield, J. F. (1991). Models of repression of transposition in P-M hybrid dysgenesis by P cytotype and by zygotically encoded repressor proteins. *Genetics, 128*(2), 471-486

Brookfield, J. F. (2005). The ecology of the genome - mobile DNA elements and their hosts. *Nat Rev Genet, 6*(2), 128-136

Burt, A., and Trivers, R. (2006). Genes in Conflict: The biology of selfish genetic elements. *Harvard University Press*

Bushnell, B. (2014) *BBMap: A Fast, Accurate, Splice-Aware Aligner*.

Canapa, A., Barucca, M., Biscotti, M. A., Forconi, M., and Olmo, E. (2015). Transposons, genome size, and evolutionary insights in animals. *Cytogenet Genome Res, 147*(4), 217-239

Carmell, M. A., Xuan, Z., Zhang, M. Q., and Hannon, G. J. (2002). The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes Dev, 16*(21), 2733-2742

Carmona, L. M., and Schatz, D. G. (2017). New insights into the evolutionary origins of the recombination-activating gene proteins and V(D)J recombination. *Febs j, 284*(11), 1590-1605

Carr, M., Bensasson, D., and Bergman, C. M. (2012). Evolutionary genomics of transposable elements in *Saccharomyces cerevisiae*. *PLoS One, 7*(11), e50978

Chuong, E. B., Elde, N. C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet, 18*(2), 71-86

Coffin, J. M., Hughes, S. H., and Varmus, H. E. (1997) *Retroviruses*. Cold Spring Harbor (NY)

Cohen, C. J., Lock, W. M., and Mager, D. L. (2009). Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene, 448*(2), 105-114

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradham, S., Stanley, F. N., Pellegrini, M., and Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature, 452*(7184), 215-219

Cooley, L., Kelley, R., and Spradling, A. (1988). Insertional mutagenesis of the *Drosophila* genome with single P elements. *Science, 239*(4844), 1121-1128

Cornelis, G., Vernochet, C., Malicorne, S., Souquere, S., Tzika, A. C., Goodman, S. M., Catzeflis, F., Robinson, T. J., Milinkovitch, M. C., Pierron, G., Hiedmann, O., Dupressoir, A., and Heidmann, T. (2014). Retroviral envelope syncytin capture in an ancestrally diverged mammalian clade for placentation in the primitive Afrotherian tenrecs. *Proc Natl Acad Sci U S A, 111*(41), E4332-4341

Craig, N. L., Chandler, M., Gellert, M., Lambowitz, A. M., Rice, P. A., and Sandmeyer, S. B. (2015). Mobile DNA III.

Czech, B., and Hannon, G. J. (2016). One loop to rule them all: the ping-pong cycle and piRNA-guided silencing. *Trends Biochem Sci, 41*(4), 324-337

de Koning, A. P., Gu, W., Castoe, T. A., Batzer, M. A., and Pollock, D. D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet, 7*(12), e1002384

de Vanssay, A., Bouge, A. L., Boivin, A., Hermant, C., Teysset, L., Delmarre, V., Antoniewski, C., and Ronsseray, S. (2012). Paramutation in *Drosophila* linked to emergence of a piRNA-producing locus. *Nature, 490*(7418), 112-115

Desset, S., Conte, C., Dimitri, P., Calco, V., Dastugue, B., and Vaury, C. (1999). Mobilization of two retroelements, *ZAM* and *Idefix*, in a novel unstable line of *Drosophila melanogaster*. *Mol Biol Evol, 16*(1), 54-66

Desset, S., Meignin, C., Dastugue, B., and Vaury, C. (2003). *COM*, a heterochromatic locus governing the control of independent endogenous retroviruses from *Drosophila melanogaster*. *Genetics, 164*(2), 501-509

Dorogova, N. V., Bolobolova, E. U., and Zakharenko, L. P. (2017). Cellular aspects of gonadal atrophy in *Drosophila* P-M hybrid dysgenesis. *Dev Biol, 424*(2), 105-112

Dunwell, T. L., McGuffin, L. J., Dunwell, J. M., and Pfeifer, G. P. (2013). The mysterious presence of a 5-methylcytosine oxidase in the *Drosophila* genome: possible explanations. *Cell Cycle, 12*(21), 3357-3365

Eggleston, W. B., Johnson-Schlitz, D. M., and Engels, W. R. (1988). P-M hybrid dysgenesis does not mobilize other transposable element families in *D. melanogaster*. *Nature, 331*(6154), 368-370

Elbarbary, R. A., Lucas, B. A., and Maquat, L. E. (2016). Retrotransposons as regulators of gene expression. *Science, 351*(6274), 7247

Engels, W. R. (1992). The origin of P elements in *Drosophila melanogaster*. *Bioessays, 14*(10), 681-686

Engels, W. R., Johnson-Schlitz, D. M., Eggleston, W. B., and Sved, J. (1990). High-frequency P element loss in *Drosophila* is homolog dependent. *Cell, Vol. 62, 515425*

Engels, W. R., and Preston, C. R. (1979). Hybrid Dysgenesis in *Drosophila melanogaster*: the biology of female and male sterility. *Genetics, 92*(1), 161-174

Esnault, C., Cornelis, G., Heidmann, O., and Heidmann, T. (2013). Differential evolutionary fate of an ancestral primate endogenous retrovirus envelope gene, the EnvV syncytin, captured for a function in placentation. *PLoS Genet, 9*(3), e1003400

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat Rev Genet, 9*(5), 397-405

Feschotte, C., and Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet, 41*, 331-368

Fukui, T., Inoue, Y., Yamaguchi, M., and Itoh, M. (2008). Genomic P elements content of a wild M' strain of *Drosophila melanogaster*: KP elements do not always function as type II repressor elements. *Genes Genet Syst, 83*(1), 67-75

Gloor, G. B., Preston, C. R., Johnson-Schlitz, D. M., Nassif, N. A., Phillis, R. W., Benz, W. K., Robertson, H. M., and Engels, W. R. (1993). Type I Repressors of P Element Mobility. *Genetics* 135, 81-95

Goriaux, C., Theron, E., Brasset, E., and Vaury, C. (2014). History of the discovery of a master locus producing piRNAs: the *flamenco*/*COM* locus in *Drosophila melanogaster*. *Front Genet, 5*, 257

Gregory, T. R. (2005). Synergy between sequence and size in large-scale genomics. *Nat Rev Genet, 6*(9), 699-708

Guio, L., Barron, M. G., and Gonzalez, J. (2014). The transposable element *Bari-Jheh* mediates oxidative stress response in *Drosophila*. *Mol Ecol, 23*(8), 2020-2030

Hagemann, A. T., and Craig, N. L. (1993). *Tn7* transposition creates a hotspot for homologous recombination at the transposon donor site. *Genetics, 133*(1), 9-16

Hamilton, A., Voinnet, O., Chappell, L., and Baulcombe, D. (2002). Two classes of short interfering RNA in RNA silencing. *EMBO J, 21*(17), 4671-4679

Hattori, M., Fujiyama, A., Taylor, T. D., Watanabe, H., Yada, T., Park, H.-S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D. K., Groner, Y., Soeda, E., Ohki, M., Takagi, T., Sakaki, Y., Taudien, S., Blechschmidt, K., Polley, A., Menzel, U., Delabar, J., Kumpf, K., Lehmann, R., Patterson, D., Reichwald, K., Rump, A., Schillhabel, M., Schudy, A., Zimmermann, W., Rosenthal, A., Kudoh, J., Schibuya, K., Kawasak,i K., Asakawa, S., Shintani, A., Sasaki, T., Nagamine, K., Mitsuyama, S., Antonarakis, S. E., Minoshima, S., Shimizu, N., Nordsiek, G., Hornischer, K., Brant, P., Scharfe, M., Schon, O., Desario, A., Reichelt, J., Kauer, G., Blocker, H., Ramser, J., Beck, A., Klages, S., Hennig, S., Riesselmann, L., Dagand, E., Haaf, T., Wehrmeyer, S., Borzym, K., Gardiner, K., Nizetic, D., Francis, F., Lehrach, H., Reinhardt, R., and Yaspo, M.-L. (2000). The DNA sequence of human chromosome 21. *Nature, 405*(6784), 311-319

Havecker, E. R., Gao, X., and Voytas, D. F. (2004). The diversity of LTR retrotransposons. In *Genome Biol* 5, 225

Hickman, A. B., and Dyda, F. (2015). The casposon-encoded Cas1 protein from *Aciduliprofundum boonei* is a DNA integrase that generates target site duplications. *Nucleic Acids Res, 43*(22), 10576-10587

Hill, T., Schlotterer, C., and Betancourt, A. J. (2016). Hybrid dysgenesis in *Drosophila simulans* associated with a rapid invasion of the P-element. *PLoS Genet, 12*(3), e1005920

Hiraizumi, Y. (1971). Spontaneous recombination in *Drosophila melanogaster* males. *Proc Natl Acad Sci U S A, 68*(2), 268-270

Hiraizumi, Y., Slatko, B., Langley, C., and Nill, A. (1973). Recombination in *Drosophila melanogaster* male. *Genetics, 73*(3), 439-444

Houwing, S., Kamminga, L. M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., Filippov, D. V., Blaser, H., Raz, E., Moens, C. B., Plasterk, R. H., Hannon, G. J., Draper, B. W., and Ketting, R. F. (2007). A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell, 129*(1), 69-82

Hua-Van, A., Le Rouzic, A., Boutin, T. S., Filee, J., and Capy, P. (2011). The struggle for life of the genome's selfish architects. *Biol Direct, 6*, 19

Huang, S., Tao, X., Yuan, S., Zhang, Y., Li, P., Beilinson, H. A., Zhang, Y., Yu, W., Pontarotti, P., Escriva, H., Le Petillon, Y., Liu, X., Chen, S., Schatz, D. G., and Xu, A. (2016). Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. *Cell, 166*(1), 102-114

Huang, X., Fejes Toth, K., and Aravin, A. A. (2017). piRNA Biogenesis in *Drosophila melanogaster*. *Trends Genet, 33*(11), 882-894

Hummel, T., and Klambt, C. (2008). P-element mutagenesis. *Methods Mol Biol, 420*, 97-117

Itoh, M., Fukui, T., Kitamura, M., Uenoyama, T., Watada, M., and Yamaguchi, M. (2004). Phenotypic stability of the P-M system in wild populations of *Drosophila melanogaster*. *Genes Genet Syst, 79*(1), 9-18

Itoh, M., Sasai, N., Inoue, Y., and Watada, M. (2001). P elements and P-M characteristics in natural populations of *Drosophila melanogaster* in the southernmost islands. *Heredity, 86 (2001) 206-212*.

Iwasaki, Y. W., Siomi, M. C., and Siomi, H. (2015). PIWI-Interacting RNA: its biogenesis and functions. *Annu Rev Biochem, 84*, 405-433

Jangam, D., Feschotte, C., and Betran, E. (2017). Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet, 33*(11), 817-831

Josse, T., Teysset, L., Todeschini, A. L., Sidor, C. M., Anxolabehere, D., and Ronsseray, S. (2007). Telomeric trans-silencing: an epigenetic repression combining RNA silencing and heterochromatin formation. *PLoS Genet, 3*(9), 1633-1643

Kapitonov, V. V., and Jurka, J. (2001). Rolling-circle transposons in eukaryotes. In *Proc Natl Acad Sci U S A* 98, 8714-8719

Kapitonov, V. V., and Jurka, J. (2005). RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. *PLoS Biol, 3*(6), e181

Kapitonov, V. V., and Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A, 103*(12), 4540-4545

Kapitonov, V. V., and Jurka, J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet, 23*(10), 521-529

Kapusta, A., and Feschotte, C. (2014). Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet, 30*(10), 439-452

Kelleher, E. S., Jaweria, J., Akoma, U., Ortega, L., and Tang, W. (2018). QTL mapping of natural variation reveals that the developmental regulator *bruno* reduces tolerance to P-element transposition in the *Drosophila* female germline. *PLoS Biol, 16*(10)

Khurana, J. S., Wang, J., Xu, J., Koppetsch, B. S., Thomson, T. C., Nowosielska, A., Theurkauf, W. E. (2011). Adaptation to P element transposon invasion in *Drosophila melanogaster*. *Cell, 147*(7), 1551-1563

Kidwell, M. G. (1985). Hybrid dysgenesis in *Drosophila melanogaster*: nature and inheritance of P element regulation. *Genetics, 111*(2), 337-350

Kidwell, M. G., and Kidwell, J. F. (1975). Cytoplasm-chromosome interactions in *Drosophila melanogaster*. *Nature, 253*(5494), 755-756

Kidwell, M. G., Kidwell, J. F., and Sved, J. A. (1977). Hybrid dysgenesis in *Drosophila melanogaster*: A syndrome of aberrant traits including mutation, sterility and male recombination. *Genetics, 86*(4), 813-833

Kidwell, M. G., and Novy, J. B. (1979). Hybrid dysgenesis in *Drosophila melanogaster*: sterility resulting from gonadal dysgenesis in the P-M System. *Genetics, 92*(4), 1127-1140

Kitano, H. (2004). Biological robustness. *Nat Rev Genet, 5*(11), 826-837

Klenov, M. S., Lavrov, S. A., Korbut, A. P., Stolyarenko, A. D., Yakushev, E. Y., Reuter, M., Pillai, R. S., Gvozdev, V. A. (2014). Impact of nuclear Piwi elimination on chromatin state in *Drosophila melanogaster* ovaries. *Nucleic Acids Res, 42*(10), 6208-6218

Klenov, M. S., Lavrov, S. A., Stolyarenko, A. D., Ryazansky, S. S., Aravin, A. A., Tuschl, T., and Gvozdev, V. A. (2007). Repeat-associated siRNAs cause chromatin silencing of retrotransposons in the *Drosophila melanogaster* germline. *Nucleic Acids Res, 35*(16), 5430-5438

Klenov, M. S., Sokolova, O. A., Yakushev, E. Y., Stolyarenko, A. D., Mikhaleva, E. A., Lavrov, S. A., and Gvozdev, V. A. (2011). Separation of stem cell maintenance and transposon silencing functions of Piwi protein. *Proc Natl Acad Sci U S A, 108*(46), 18760-18765

Kofler, R., Hill, T., Nolte, V., Betancourt, A. J., and Schlötterer, C. (2015). The recent invasion of natural *Drosophila simulans* populations by the P-element. In *Proc Natl Acad Sci U S A* 112, 6659-6663

Kofler, R., Senti, K. A., Nolte, V., Tobler, R., and Schlötterer, C. (2018). Molecular dissection of a natural transposable element invasion. *Genome Res, 28*(6), 824-835

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res*, 27(5):722-736

Kronmiller, B. A., and Wise, R. P. (2008). TEnest: automated chronological annotation and visualization of nested plant transposable elements. *Plant Physiol, 146*(1), 45-59

Krupovic, M., and Makarova, K. S. (2014). *Casposons* a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biol* 19, 12-36

Kuramochi-Miyagawa, S., Watanabe, T., Gotoh, K., Totoki, Y., Toyoda, A., Ikawa, M., Asada, N., Kojima, K., Yamaguchi, Y., Ijiri, T. W., Hata, K., Li, E., Matsuda, Y., Kimura, T., Okabe, M., Sakaki, Y., Sasaki, H., and Nakano, T. (2008). DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes Dev, 22*(7), 908-917

Lachaise, D., David, J. R., Lemeunier, F., Tsacas, L., and Ashburner, M. (1986). The reproductive relationships of *Drosophila sechellia* with *D. maurittiana , D. simulans*, and *D. melanogaster* from the afrotropical region. *Evolution, 40*(2), 262-271

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S, Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs,

R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R, Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A. P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A, Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and Chen, Y. J.

(2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921

Langley, C. H., Montgomery, E., Hudson, R., Kaplan, N., and Charlesworth, B. (1988). On the role of unequal exchange in the containment of transposable element copy number. *Genet Res, 52*(3), 223-235

Laski, F. A., Rio, D. C., and Rubin, G. M. (1986). Tissue specificity of *Drosophila* P element transposition is regulated at the level of mRNA splicing. *Cell, 44*(1), 7-19

Lavialle, C., Cornelis, G., Dupressoir, A., Esnault, C., Heidmann, O., Vernochet, C., and Heidmann, T. (2013). Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. In *Philos Trans R Soc Lond B Biol Sci,* 368(1626): 20120507

Le Thomas, A., Rogers, A. K., Webster, A., Marinov, G. K., Liao, S. E., Perkins, E. M., Hur, E. K., Aravin, A. A., and Toth, K. F. (2013). Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes Dev, 27*(4), 390-399

Le Thomas, A., Toth, K. F., and Aravin, A. A. (2014). To be or not to be a piRNA: genomic origin and processing of piRNAs. *Genome Biol, 15*(1), 204

Lee, C. C., Mul, Y. M., and Rio, D. C. (1996). The *Drosophila* P-element KP repressor protein dimerizes and interacts with multiple sites on P-element DNA. *Mol Cell Biol, 16*(10), 5616-5622

Lee, Y. C., and Karpen, G. H. (2017). Pervasive epigenetic effects of *Drosophila* euchromatic transposable elements impact their evolution. *Elife, 6*

Lee, Y. C., and Langley, C. H. (2012). Long-term and short-term evolutionary impacts of transposable elements on *Drosophila*. *Genetics, 192*(4), 1411-1432

Lemaitre, B., Ronsseray, S., and Coen, D. (1993). Maternal repression of the P element promoter in the germline of *Drosophila melanogaster*: a model for the P cytotype. *Genetics, 135*(1), 149-160

Lewis, S. H., Quarles, K. A., Yang, Y., Tanguy, M., Frezal, L., Smith, S. A., Sharma, P. P., Cordaux, R., Gilbert. C., Giraud, I., Collins, D.H., Zamore P.D., Miska E.A., Sarkies, P., and Jiggins, F. M. (2018). Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. *Nat Ecol Evol, 2*(1), 174-181

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079

Liu, R., Holik, A. Z., Su, S., Jansz, N., Chen, K., Leong, H. S., Blewitt, M. E., Asselin-Labat, M. L., Smyth, G. K., and Ritchie, M. E. (2015). Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res, 43*(15), e97

Luan, D. D., Korman, M. H., Jakubczak, J. L., and Eickbush, T. H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell, 72*(4), 595-605

Malone, C. D., Brennecke, J., Dus, M., Stark, A., McCombie, W. R., Sachidanandam, R., and Hannon, G. J. (2009). Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell, 137*(3), 522-535

Marcais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol, 14*(1), e1005944

Marcel M. (2017) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet*, 17(1):10-12

Martens, J. H. A., O'Sullivan, R. J., Braunschweig, U., Opravil, S., Radolf, M., Steinlein, P., and Jenuwein, T. (2005). The profile of repeat-associated histone lysine methylation states in the mouse epigenome. In *EMBO J* 24, 800-812

McVey, M., Larocque, J. R., Adams, M. D., and Sekelsky, J. J. (2004). Formation of deletions during double-strand break repair in *Drosophila* DmBlm mutants occurs after strand invasion. *Proc Natl Acad Sci U S A, 101*(44), 15694-15699

Mendiola, M. V., Bernales, I., and de la Cruz, F. (1994). Differential roles of the transposon termini in IS91 transposition. *Proc Natl Acad Sci U S A, 91*(5), 1922-1926

Miller, W. J., McDonald, J. F., and Pinsker, W. (1997). Molecular domestication of mobile elements. *Genetica, 100*(1-3), 261-270

Misra, S., and Rio, D. C. (1990). Cytotype control of *Drosophila* P element transposition: the 66 kd protein is a repressor of transposase activity. *Cell, 62*(2), 269-284

Mohn, F., Sienski, G., Handler, D., and Brennecke, J. (2014). The rhino-deadlock-cutoff complex licenses noncanonical transcription of dual-strand piRNA clusters in *Drosophila*. *Cell, 157*(6), 1364-1379

Montgomery, E., Charlesworth, B., and Langley, C. H. (1987). A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. *Genet Res, 49*(1), 31-41

Moon, S., Cassani, M., Lin, Y. A., Wang, L., Dou, K., and Zhang, Z. Z. (2018). A robust transposon-endogenizing response from germline stem cells. *Dev Cell, 47*(5), 660-671.e663

Nassif, N., Penney, J., Pal, S., Engels, W. R., and Gloor, G. B. (1994). Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. *Mol Cell Biol, 14*(3), 1613-1625

Nattestad, M., and Schatz, M. C. (2016). Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics, 32*(19), 3021-3023

Noutsopoulos, D., Markopoulos, G., Vartholomatos, G., Kolettas, E., Kolaitis, N., and Tzavaras, T. (2010). VL30 retrotransposition signals activation of a caspase-independent and p53-dependent death pathway associated with mitochondrial and lysosomal damage. *Cell Res, 20*(5), 553-562

O'Hare, K., and Rubin, G. M. (1983). Structures of P transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. *Cell, 34*(1), 25-35

Ogura, K., Woodruff, R. C., Itoh, M., and Boussy, I. A. (2007). Long-term patterns of genomic P element content and P-M characteristics of *Drosophila melanogaster* in eastern Australia. *Genes Genet Syst, 82*(6), 479-487

Olovnikov, I., Ryazansky, S., Shpiz, S., Lavrov, S., Abramov, Y., Vaury, C., Jensen, S., and Kalmykova, A. (2013). De novo piRNA cluster formation in the *Drosophila* germ line triggered by transgenes containing a transcribed transposon fragment. In *Nucleic Acids Res* 41, 5757-5768

Orgel, L. E., and Crick, F. H. (1980). Selfish DNA: the ultimate parasite. *Nature, 284*(5757), 604-607

Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D., and Zamore, P. D. (2019). PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet, 20*(2), 89-108

Pace, J. K., 2nd, and Feschotte, C. (2007). The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res, 17*(4), 422-432

Parisi, M. J., Deng, W., Wang, Z., and Lin, H. (2001). The arrest gene is required for germline cyst formation during *Drosophila* oogenesis. *Genesis, 29*(4), 196-209

Pasyukova, E. G., Nuzhdin, S. V., Morozova, T. V., and Mackay, T. F. (2004). Accumulation of transposable elements in the genome of *Drosophila melanogaster* is associated with a decrease in fitness. *J Hered, 95*(4), 284-290

Pelisson, A., Song, S. U., Prud'homme, N., Smith, P. A., Bucheton, A., and Corces, V. G. (1994). *Gypsy* transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the *Drosophila flamenco* gene. *EMBO J, 13*(18), 4401-4411

Piskurek, O., and Jackson, D. J. (2012). Transposable elements: from DNA parasites to architects of metazoan evolution. *Genes (Basel)* 3, 409-422

Plasterk, R. H. (1991). The origin of footprints of the Tc1 transposon of *Caenorhabditis elegans*. *EMBO J, 10*(7), 1919-1925

Plasterk, R. H., and Groenen, J. T. (1992). Targeted alterations of the *Caenorhabditis elegans* genome by transgene instructed DNA double strand break repair following Tc1 excision. *EMBO J, 11*(1), 287-290

Pritham, E. J., Putliwala, T., and Feschotte, C. (2007). *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene, 390*(1-2), 3-17

Qi, Y., He, X., Wang, X. J., Kohany, O., Jurka, J., and Hannon, G. J. (2006). Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature, 443*(7114), 1008-1012

Rangan, P., Malone, C. D., Navarro, C., Newbold, S. P., Hayes, P. S., Sachidanandam, R., Lehmann, R. (2011). piRNA production requires heterochromatin formation in *Drosophila*. *Curr Biol, 21*(16), 1373-1379

Rebollo, R., Romanish, M. T., and Mager, D. L. (2012). Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet, 46*, 21-42

Rio, D. C., Laski, F. A., and Rubin, G. M. (1986). Identification and immunochemical analysis of biologically active *Drosophila* P element transposase. *Cell, 44*(1), 21-32

Ronsseray, S. (2015). Paramutation phenomena in non-vertebrate animals. *Semin Cell Dev Biol, 44*, 39-46

Ronsseray, S., Lehmann, M., and Anxolabehere, D. (1991). The maternally inherited regulation of P elements in *Drosophila melanogaster* can be elicited by two P copies at cytological site 1A on the X chromosome. *Genetics, 129*(2), 501-512

Rozhkov, N. V., Hammell, M., and Hannon, G. J. (2013). Multiple roles for Piwi in silencing *Drosophila* transposons. *Genes Dev, 27*(4), 400-412

Schaefer, R. E., Kidwell, M. G., and Fausto-Sterling, A. (1979). Hybrid dysgenesis in *Drosohila melanogaster*: morphological and cytological studies of ovarian dysgenesis. *Genetics, 92*(4), 1141-1152

Senti, K. A., and Brennecke, J. (2010). The piRNA pathway: a fly's perspective on the guardian of the genome. *Trends Genet, 26*(12), 499-509

Sienski, G., Donertas, D., and Brennecke, J. (2012). Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell, 151*(5), 964-980

Slotkin, R. K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet, 8*(4), 272-285

Smit, A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev, 9*(6), 657-663

Srivastav, S. P. and Kelleher, E. S. (2017). Paternal induction of hybrid dysgenesis in *Drosophila melanogaster* is weakly correlated with both P-element and *hobo* element dosage. *G3, 7*(5), 1487-1497

Sturtevant, A. H. (1920). Genetic Studies on *Drosophila simulans*. I. Introduction. Hybrids with *Drosophila melanogaster*. *Genetics, 5*(5), 488-500

Svetec, N., Cridland, J. M., Zhao, L., and Begun, D. J. (2016). The adaptive significance of natural genetic variation in the DNA damage response of *Drosophila melanogaster*. *PLoS Genet, 12*(3), e1005869

Tasnim, S., and Kelleher, E. S. (2018). p53 is required for female germline stem cell maintenance in P-element hybrid dysgenesis. *Dev Biol, 434*(2), 215-220

Teixeira, F. K., Okuniewska, M., Malone, C. D., Coux, R. X., Rio, D. C., and Lehmann, R. (2017). piRNA-mediated regulation of transposon alternative splicing in the soma and germ line. *Nature, 552*(7684), 268-272

Thompson, P. J., Macfarlan, T. S., and Lorincz, M. C. (2016). Long Terminal Repeats: from parasitic elements to building blocks of the transcriptional regulatory repertoire. *Mol Cell, 62*(5), 766-776

Todeschini, A. L., Teysset, L., Delmarre, V., and Ronsseray, S. (2010). The epigenetic trans-silencing effect in *Drosophila* involves maternally-transmitted small RNAs whose production depends on the piRNA pathway and HP1. *PLoS One, 5*(6), e11032

Vagin, V. V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., and Zamore, P. D. (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. *Science, 313*(5785), 320-324

Wacholder, A. C., Cox, C., Meyer, T. J., Ruggiero, R. P., Vemulapalli, V., Damert, A., Pollock, D. D. (2014). Inference of transposable element ancestry. *PLoS Genet, 10*(8)

Wakisaka, K. T., Ichiyanagi, K., Ohno, S., and Itoh, M. (2017). Diversity of P-element piRNA production among M' and Q strains and its association with P-M hybrid dysgenesis in *Drosophila melanogaster*. *Mob DNA, 8*, 13

Wakisaka, K. T., Ichiyanagi, K., Ohno, S., and Itoh, M. (2018). Association of zygotic piRNAs derived from paternal P elements with hybrid dysgenesis in *Drosophila melanogaster*. *Mob DNA, 9*, 7

Wang, S. H., and Elgin, S. C. (2011). *Drosophila* Piwi functions downstream of piRNA production mediating a chromatin-based transposon silencing mechanism in female germ line. *Proc Natl Acad Sci U S A, 108*(52), 21164-21169

Wang, Z., and Lin, H. (2007). Sex-lethal is a target of *bruno*-mediated translational repression in promoting the differentiation of stem cell progeny during *Drosophila* oogenesis. *Dev Biol, 302*(1), 160-168

Werren, J. H., Nur, U., and Wu, C. I. (1988). Selfish genetic elements. *Trends Ecol Evol* 3, 297-302

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., . . . Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet, 8*(12), 973-982

Yamanaka, S., Siomi, M. C., and Siomi, H. (2014). piRNA clusters and open chromatin structure. In *Mob DNA* 5, 22

Yannopoulos, G., Stamatis, N., Monastirioti, M., Hatzopoulos, P., and Louis, C. (1987). *hobo* is responsible for the induction of hybrid dysgenesis by strains of *Drosophila melanogaster* bearing the male recombination factor 23.5MRF. *Cell, 49*(4), 487-495

Zanni, V., Eymery, A., Coiffet, M., Zytnicki, M., Luyten, I., Quesneville, H., Vaury, C., and Jensen, S. (2013). Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proc Natl Acad Sci U S A, 110*(49), 19842-19847

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W., Chen, H., Henderson, I. R.,Shinn, P., and Ecker, J. R. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell, 126*(6), 1189-1201

Zhang, Z., Wang, J., Schultz, N., Zhang, F., Parhad, S. S., Tu, S., Vreven, T., Zamore, P.D., Weng, Z., and Theurkauf, W. E. (2014). The HP1 homolog rhino anchors a nuclear complex that suppresses piRNA precursor splicing. *Cell, 157*(6), 1353-1363